WAC 2026

The 13th Edition of the Web-as-Corpus Workshop

Proceedings of the Workshop

October 29, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-481-1

Introduction

The Web as Corpus (WaC) workshop series has a long history. The first WaC workshop took place at the Corpus Linguistics conference in Birmingham in 2005, at a time when treating the Web as a serious source of linguistic data was still far from standard practice. The attraction was obvious: the Web offered quantities and varieties of language data that traditional corpus construction could hardly match, particularly for less-resourced languages. At the same time, using it as a corpus meant confronting difficult questions of sampling, representativeness, language identification, cleaning, duplication, metadata, and the very meaning of a corpus when its source was something as large, dynamic, and heterogeneous as the Web.

WaC-13 marks a return of the workshop after a substantial hiatus. WaC-XI took place in 2017, while WaC-XII, planned for LREC 2020, had its physical meeting cancelled by the COVID-19 pandemic, although its proceedings were published. In the intervening years, the role of web data in language technology has changed dramatically. With the advent of large language models, web corpora have moved from being a sometimes unconventional source of language data to becoming a central component of the data infrastructure underlying modern NLP. Questions that occupied the WaC community from its beginnings now concern a much larger part of our field. What does a web crawl contain? What does it miss? How should it be filtered? Which languages, registers, communities, and kinds of content survive the successive stages from crawling to a released corpus? And what biases are introduced along the way?

The programme of WaC-13 reflects this new centrality of web data, while at the same time retaining the remarkable breadth that has always characterized Web-as-Corpus research. Several contributions examine the fundamental processing stages of contemporary web corpora: language identification, main-content extraction, quality filtering, register classification, and deduplication. Others ask what our pipelines systematically lose, with studies of endangered and low-resource languages, multilingual coverage, and the effects of filtering on geographical, sociolinguistic, thematic, and register diversity. Questions of representativeness span languages and regions from Sámi and Muskogean to Persian, South Slavic, Finnish, Sri Lankan, and cross-lingual resources.

At the same time, the programme shows how the object of study itself is changing. Contributions investigate AI-generated text on the Web, AI-mediated search, machine-translated web corpora, changing patterns in the questions found online, and the emerging infrastructure of crawler consent through `robots.txt`. Other papers extend the notion of Web as Corpus beyond conventional collections of HTML text, considering PDFs, web video and speech, online debates, Wiktionary, news, and even programming languages. Taken together, the papers remind us that there is no single "web corpus": there are successive choices about what to crawl, extract, identify, retain, annotate, and ultimately make available for research and model training.

Given the changes the field has undergone in recent years, it is particularly fitting that our invited speaker is Sebastian Nagel of the Common Crawl Foundation. His talk, *The Web as a Corpus and Common Crawl: A Look Back, Forward and into Present Challenges*, connects the history of the field directly to its current infrastructure. Common Crawl occupies a special position in today's web-data ecosystem. By maintaining an openly accessible, continuously updated archive of web crawl data, it provides infrastructure on which a substantial amount of academic and industrial research now depends. For researchers without the resources to operate a web-scale crawler themselves, such open infrastructure is especially important: access to the raw material of web-scale research should not be restricted to a small number of very large organizations. The prominence of Common Crawl in the WaC-13 programme itself, both as a source of data and as an object of critical investigation, also demonstrates that shared infrastructure is most valuable when a research community can study its strengths, limitations, coverage, and consequences openly.

Finally, WaC-13 would not have happened without SIGWaC, the ACL Special Interest Group on Web as Corpus. There is an appealing continuity here with the history of the workshop. Web-as-Corpus research grew to a large extent from the bottom up: researchers discovered that the Web offered extraordinary possibilities, developed methods and tools to make use of it, and formed a community around the

problems they encountered. SIGWaC, and the broader framework of the Association for Computational Linguistics, complement this bottom-up energy with something equally important: durable structures through which a research community can maintain contact, exchange knowledge, organize events, and renew itself across generations of researchers. Such structures should not determine where research goes: the most interesting developments will continue to emerge from researchers pursuing questions that were not anticipated by any committee or organization. That said, conferences, workshops, and SIGs give those developments places in which to meet. They help turn scattered individual efforts into conversations, and conversations into communities.

After a long break, we are particularly happy to bring the Web as Corpus community together again at WaC-13. We would like to thank all authors who submitted to the workshop, the reviewers who gave their time to provide valuable feedback on the submissions, and the EMNLP workshop chairs for their support. The technological context may have changed profoundly since 2005, but the central insight behind the workshop has proved remarkably durable: the Web is an extraordinary source of evidence about language, but turning the Web into data is never a neutral or trivial operation. In the age of large language models, understanding that process matters more than ever.

WaC-13 Organizing Committee

Organizing Committee

Program Chairs

Pedro Ortiz Suarez, Common Crawl Foundation, USA
Laurie Burchell, Common Crawl Foundation, USA
Nikola Ljubešić, Jožef Stefan Institute, Slovenia
Yves Scherrer, University of Oslo, Norway
Veronika Laippala, University of Turku, Finland
Vuk Dinić, Jožef Stefan Institute, Slovenia
Thom Vaughan, Common Crawl Foundation, USA

Program Committee

Program Chairs

Laurie Burchell, Common Crawl Foundation
Vuk Dinić, Jozef Stefan Institute and Jožef Stefan International Postgraduate School
Veronika Laippala, University of Turku
Nikola Ljubešić, Jožef Stefan Institute
Pedro Ortiz Suarez, Common Crawl Foundation
Yves Scherrer, University of Oslo
Thom Vaughan, Common Crawl Foundation

Reviewers

Mantaqa Abedin, Dominik Allen, Catherine Arnett

Pushkar Baranwal, Vladimír Benko, Arghya Bhattacharya, Magnus Breder Birkenes, Laurie Burchell, Albina Burlova

Hande Celikkanat, Aditya Channa, Antonis Charalampous, Pinzhen Chen, Austin Cheng, Oleg Chernikov, Steven Coats, Sophia Conrad

Ahrnish P Dahal, Constantinos Djouvas, Ona de Gibert

Valeria Efimova

Adriano Ferraresi, Marie Flesch, Luca Foppiano

Tomasz Galica, Susmita Gangopadhyay, Filip Ginter

Srinivas Harish, Jindřich Helcl, Erik Henriksson, Im Hyeonyeol

Maximilian Idahl

Sangwan Jeon, Robert Jäschke

Madhav Kapoor, Toprak Seda Karaosmanoğlu, Parul Khandelwal, Mateusz Klimaszewski, Jaroslav Kopcan, Olli Kuparinen, Andrey Kutuzov

Veronika Laippala, Yibo Li, Geng Liu, Nikola Ljubešić

Vukosi Marivate, Bruno Leonardo Santos Menezes

Kumarmanas Nethil

Stephan Oepen, Pedro Ortiz Suarez, Malte Ostendorff, Christian Oswald

Michael Paris, Charles Pert, Francesco Pierri, Paola Pietrandrea

Nishat Raihan, Tharindu Ranasinghe, Rishi Ravikumar, Georg Rehm, Jacob Rigal, Daniel Ruffinelli

Kengatharaiyer Sarveswaran, Nadav Shani, Serge Sharoff, Jastej Singh, Preety Singh, Rob Sneiderman, Amr Sobhy, Egon Stemle, Vít Suchomel, Marek Šuppa

Sotaro Takeshita, Otto Tarkka, Lin Tian, Luisa Troncone

Sami Virpioja, Jelmer van der Linde

Jessica Wang, Sidney Wong

Marcos Zampieri, Gary Zhang, Ruiyang Zhang, Calvin Zhou

Keynote Talk

The Web as a Corpus and Common Crawl: A look back, forward and into present challenges

Sebastian Nagel
Common Crawl Foundation



2026-10-29 09:15:00 – Room: **Room 1**

Abstract: During the 1990s, the use of text corpora to study natural language became widespread. This shift away from intuition to data-driven methodologies based on representative collections of language in use became known as the "corpus revolution". Corpora grew in size, reaching the landmark figure of 100 million words. Corpora also became digital during this decade.

During the same decade, the World Wide Web, invented in 1989, saw a swift breakthrough and began its exponential growth. It was therefore only natural that corpus linguists started to use the web as a resource. The "Web as a Corpus" emerged, and around 2005, the SIGWAC and WaCky initiatives started their work. Using web-harvested texts in the same way as established corpora came with many challenges, including selecting a suitable set of seed URLs, implementing and running a web crawler, extracting text from HTML, cleaning boilerplate content, detecting duplicates and near-duplicates, adding linguistic annotations and building user-friendly interfaces that enable linguists to explore the corpora.

Another initiative, Common Crawl, was founded in 2007. Its mission is to make web data more accessible to programmers and data scientists. Rather than operating a web crawler themselves, developers and researchers can focus on using web data to gain insights and build innovative products. Common Crawl published its first data set in 2008. Since then, more than 120 large-scale web crawls have been run, and the data has been used and cited in over 10,000 research papers. While a significant proportion of this research is about NLP and language modelling, Common Crawl's scope extends to many other disciplines, including web science, information retrieval, the semantic web, internet security and digital forensics, and humanities and social sciences.

Around 2010, both the Web-as-Corpus initiative and Common Crawl saw a surge in interest and growing communities. A few years later, however, interest dropped. Although many tools for processing noisy web data were already available, scaling up and tuning these tools for a particular use case or language remained challenging. It was not until 2020, with the advent of the transformer era in language modelling coupled with the general availability of cloud computing, that it became possible to process terabytes and petabytes of web data. Since then, training language models has become the most important use case for web data. Nevertheless, many of the challenges described in the early Web-as-Corpus workshops remain unsolved and require continuous, collaborative research.

This talk will look back at the early days of corpus linguistics and the use of web data. It will discuss the differences and similarities between the Web-as-Corpus initiatives and Common Crawl, and provide an overview of recent challenges and future tasks.

Bio: Sebastian Nagel is a programmer and engineer at Common Crawl and the University of Konstanz. At Common Crawl, he runs and maintains the crawler and works to make web data more accessible. Prior to joining Common Crawl, Sebastian worked as a software developer at Exorbyte, where he implemented search and data quality solutions. Sebastian holds a PhD in computational linguistics from the University of Munich. As a longtime committer of Nutch and StormCrawler, and a member of the Apache Software Foundation, he strongly believes in open-source software development. He studied linguistics (Slavic languages) and cultural anthropology in Munich, Kazan and Prague. It was in Prague that he first attended a class in corpus linguistics, which sparked his interest in computational linguistics.

Table of Contents

CHALIS: A Challenge Dataset for Language Identification in Difficult Scenarios
Michal Tichý and Jindřich Libovický 1

The CreateDebate Arguments Dataset: A Large-Scale Naturalistic Corpus for Computational Argumentation
Antonis Charalampous and Constantinos Djouvas 8

Who Published This Headline? Editorial Fingerprinting in French News Under Topic Control
Amr Sobhy 22

A Taxonomy of Programming Languages for Code Generation
Nishat Raihan, Christian Newman and Marcos Zampieri 37

What Persists across Web Crawls? Genre and News Topic Variation in South Slavic Web Corpora
Nikola Ljubešić, Vuk Dinić and Peter Rupnik 53

HPLT3-PerRef: A Refined Large-Scale Persian Web Corpus via Supervised Document Quality Filtering
Alireza Razzaghi, Tomasz Galica, Erik Henriksson and Veronika Laippala 60

Host-Aware Web Crawling for LLM Pretraining Data
Malte Ostendorff 70

Auditing Source Exposure in Baidu and Google AI Search
Yibo Li, Enci Guan, Yuedan Cai, Geng Liu and Francesco Pierri 88

Lost in Translation: Structured-Span Corruption in Machine-Translated Multilingual Web Corpora
Ruiyang Zhang 100

Counting Documents Is Not Counting Text: Unit Bias in Web-PDF Corpus Statistics
Luca Foppiano 107

Is Educational Content Enough? The Linguistic and Thematic Effects of Educational Filtering
Urho Kunnas, Veronika Laippala, Sampo Pyysalo and Tomasz Galica 115

WiNCC: A Reproducible Pipeline for Constructing a Wikipedia-enriched Nepali Cross-lingual Corpus
Ahrnish P Dahal, Pushkar Baranwal, Amit Pundir, Geetika Jain Saxena and Sanjeev Singh . . 131

Mapping Sámi Languages in a Norwegian National Web Crawl: Results from a Representativeness Study
Magnus Breder Birkenes 139

Towards More Reliable Web Metadata: Do Register Classifiers Learn Register or Genre Features?
Antti Kanner, Tuomas Lundberg, Erik Henriksson and Veronika Laippala 147

Same Pages, Cleaner Labels: Language Identification for Eight Low-Resource Languages in mC4 and FineWeb-2
Gary Zhang 161

How Many Duplicates Are Left? Capture-Recapture and Sampling Inversion for Residual Duplication in FineWeb
Gary Zhang 170

Using Search Engines to Trace Metalinguistic Discussions of the Noun "Female" in Digital Discourse
Marie Flesch 182

Keeping the Ums: Conversation-Analysis-Informed Verbatim ASR for Web Speech
Steven Coats 192

Segment-Level Register Analysis of the Multilingual Web
Erik Henriksson, Alireza Razzaghi, Tomasz Galica, Jiaqi Feng Guo, Anastasia Zhi-Ling Dong, Urho Kunnas, Erofili Psaltaki, Tuomas Lundberg, Antti Kanner and Veronika Laippala 200

Invisible on the Web: An Audit of Low-Resource Muskogean Languages in Web-Crawled Corpora
Sharvin Goyal 214

The Crawl Gap: Where Endangered-Language Web Content Is Lost Before It Reaches a Corpus
Sharvin Goyal 220

You Can Tell Who's Asking: What the Web's Questions Are Made Of, and Where They Come From
Calvin Zhou, Vincent McCloskey and Krishna Srinivasan 224

Wiktionary as a Crowdsourced Lexicon for English Dialects
Sidney Wong 238

Task-Agnostic Web Document Annotation with LLM-Generated Descriptors
Otto Tarkka, Erik Henriksson, Jenna Kanerva and Filip Ginter 252

PAWN++ in the Wild: Evaluating AI-Generated Text Detection on the Web
Oleg Chernikov, Evgenii Ptitsyn and Albina Burlova 282

Extraction Changes the Route: Main-Content Extraction and Language Identification in Web Corpora
Toprak Seda Karaosmanoğlu 297

CeyNews: A Sri Lankan Trilingual News Corpus
Ajintha Sivakulasingam, Kengatharaiyer Sarveswaran and Tharindu Ranasinghe 308

Probing Pretraining "Quality": Sociolinguistic Mobility as a Lens on Web Corpus Filters
Jessica Wang 320

A Web-Derived Typology of Multimodal Infant Crying: Taxonomy, a Text-Only Web-Corpus Comparison, and Ethical Issues
Jacob Rigal and Jonathan Ginzburg 331

Whose Robots.txt? Consent-Infrastructure Inequality Across the Languages of Web Corpora
Gary Zhang and Austin Cheng 347

CHALIS: A Challenge Dataset for Language Identification in Difficult Scenarios

Michal Tichý and Jindřich Libovický
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
V Holešovičkách 747/2, 180 00 Praha, Czechia
libovicky@ufal.mff.cuni.cz

Abstract

We present **CHALIS** (Challenging Language Identification Samples), a new benchmark dataset explicitly designed to address difficult cases in language identification: cousin languages and orthographic noise. Our dataset has two parts: First, we collected sentences shared across mutually intelligible language pairs (Czech/Slovak, Spanish/Catalan, Portuguese/Galician, Danish/Norwegian). The second part tests for orthography noise: we transliterate text across multiple scripts, remove diacritics, simulate homoglyph attacks, and use Internet slang. We evaluate four widely used language identification systems on CHALIS and demonstrate that all struggle substantially in these scenarios, especially on lower-resource languages within cousin pairs and on transliterated input. The resource is publicly available at https://huggingface.co/datasets/michal-tichy/CHALIS.

1 Introduction

Language identification (LangID), the task of determining the language of a given text, is a foundational component of many NLP pipelines (Jauhiainen et al., 2018). Although some researchers have characterized it as a solved problem (McNamee, 2005), growing demand for large-scale web-crawled multilingual corpora (Penedo et al., 2025; Burchell et al., 2025) has exposed significant weaknesses in current systems.

Kreutzer et al. (2022) audited 205 languages across five datasets and found that over 8% of samples were mislabeled with the wrong language, and a further 10% were not linguistic in character at all. The most error-prone cases involve languages that are closely related to a higher-resource neighbor, and texts that have been transliterated into a different script.

These findings motivate our targeted evaluation benchmark. Existing datasets used to assess LangID models, such as FLORES-200 (Team et al., 2022), originally designed for multilingual machine translation, do not specifically probe these hard cases.

We introduce CHALIS to fill this gap. The dataset covers both **cousin language pairs** and **transliteration and orthographic variation**, with a human-annotated subset for cousin language pairs where annotators validated whether sentences are grammatically correct in each language. We evaluate four state-of-the-art LangID systems on CHALIS, revealing systematic failure modes in both scenarios.

2 Background

2.1 Language Identification

Language identification is a classic text classification task (Jauhiainen et al., 2018). Modern systems such as FastText (Joulin et al., 2017), OpenLID (Burchell et al., 2023), GlotLID (Kargaran et al., 2023), and GCLD3 achieve strong results on standard benchmarks. However, as Caswell et al. (2020) shows, performance degrades significantly in realistic web-crawled settings, where mislabeled data is common.

2.2 Cousin Languages

Defining a "language" is a notoriously difficult exercise, nicely illustrated by the Weinreich witticism: *"a language is a dialect with an army and navy"* (Maxwell, 2018). Language communities that coexist geographically develop varieties that mutually influence each other, often remaining *mutually intelligible* (Chambers and Trudgill, 1998). Examples include Norwegian and Danish, or Czech and Slovak.

We refer to such pairs as **language cousins**. They pose a dual challenge for LangID: they are not equally represented in training data (one is typically less prevalent), and there exist sentences that

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 1–7
October 29, 2026 ©2026 Association for Computational Linguistics

are grammatically correct in *both* languages. For example, *"Je to obrovský problém"* ("It is a huge problem") is valid in both Czech and Slovak. Both Kreutzer et al. (2022) and Caswell et al. (2020) note that mistaking a lower-resource language for its more prominent cousin is among the most frequent errors in web-crawled corpora.

2.3 Non-standard Orthography

Transliteration, i.e., converting text from one script to another, often approximating pronunciation, is another major source of LangID errors. Transliterated text does not match the script profile that LangID models expect, yet it retains the language's lexical and syntactic patterns. For instance, Russian Cyrillic transliterated to Latin characters is meaningless to a model that associates Russian exclusively with the Cyrillic script. Moreover, different languages have different transliteration standards.

The most frequent orthography issues in writing are probably mistakes and typos. These naturally occur in most digital texts and presumably also in datasets used for LangID training. However, there are more systematic issues that might affect the LangID performance. These issues include removing diacritics and only using ASCII characters in Latin, historically driven by software and hardware limitations. Homoglyph attacks use the fact that some graphemes look similar across scripts (e.g., o and o in Latin and Cyrillic) and are often used to confuse automatic systems such as spam detectors.

3 The CHALIS Dataset

We collect a challenge set that addresses the issues discussed in the previous section. The dataset consists of three parts: Sentences that are valid in two cousin languages (and multiclass classifiers need to decide for one of them only), sentences that are only valid in one of the cousin-language pairs, but consist of words that are common in both languages, and the third part focuses on the orthography phenomena, and unlike the first two, it was created fully automatically.

Basic statistics of the dataset are presented in Table 1. The dataset is available at Huggingface Hub[1] under the CC-BY-NC 4.0 license, which is consistent with the underlying data. More technical details are in Appendix C.

[1] https://huggingface.co/datasets/michal-tichy/CHALIS

	Subset / Language	Lines	Avg. Words
Joint	Czech & Slovak	655	5.9
	Spanish & Catalan	642	5.8
	Danish & Norwegian	459	7.0
	Portuguese & Galician	1,143	7.9
Single	Czech	491	6.5
	Danish	361	8.7
	Norwegian	327	7.9
	Spanish	546	6.5
Ortography	Cyrilics to English	50k	15.4
	Cyrilics to Czech	10k	13.9
	Arabic (Kazakh)	10k	12.0
	Diacritics	30k	17.2
	Homoglyps	30k	14.5
	Antspeak	10k	11.8
	Leetspeak	10k	14.3

Table 1: CHALIS dataset statistics.

3.1 Cousin Language Data

We collected data for four cousin pairs, all Indo-European languages from Europe. Czech/Slovak (Slavic), Spanish/Catalan (Romance), Portuguese/Galician (Romance), Danish/Norwegian (Germanic). There is a different level of resource disparities between the languages. Czech and Slovak are both official languages of the countries where they are spoken, with roughly 10M and 5M speakers respectively. There is a similar situation with Danish and Norwegian, both of which are spoken by roughly 5M people. The situation is slightly different with the Romance languages. Due to the colonial history, Spanish is spoken by 500M speakers, whereas Catalan is spoken by only 10M people in Europe. Similarly, Portuguese is spoken by 250M people in Europe and South America, whereas Galician is only spoken by around 3M people in Spain.

Candidate sentence retrieval. For each language in a pair, we tokenized text using Moses tokenizer (Koehn et al., 2007) and built a frequency dictionary from the WMT Monolingual News Crawl (Kocmi et al., 2024), supplemented by Wikipedia-sourced data from the Leipzig Corpora Collection (Goldhahn et al., 2012) for lower-resource languages. We extracted the 10,000 most frequent words in each language and computed their intersection to obtain a shared vocabulary. Table 2 shows the resulting shared vocabulary sizes.

We then retrieved all sentences from the source corpora whose tokens all appear in the shared vocabulary, yielding candidate sentences that *could* belong to either language.

		Count
Czech	Slovak	2,339
Danish	Norwegian	4,266
Spanish	Catalan	1,623
Portuguese	Galician	3,491

Table 2: Shared vocabulary sizes per language pair (out of 10k must frequent tokens in the languages).

KenLM scoring. Candidates were scored using 4-gram KenLM language models (Heafield et al., 2013) with Kneser–Ney smoothing (Kneser and Ney, 1995), trained on the source corpora. Each sentence received a probability under each language's model, normalized for length. We ranked sentences by the harmonic mean of the two normalized log-likelihoods, and selected the top 1,200 per pair as most plausibly correct in both languages.

Named entity filtering. Inspection of high-scoring sentences revealed that many consisted largely of proper names (e.g., sports results). We applied the GLiNER named-entity recognizer (Zaratiana et al., 2024) to remove sentences in which more than 25% of tokens were classified as named entities (persons, locations, organizations, etc.). Near-duplicate sentences were also removed using an edit-distance threshold of 10% of sentence length.

Human annotation. The top 1,200 filtered sentences per pair were presented to human annotators, who were native speakers or advanced students of the target languages. Annotators labeled each sentence as valid or invalid in their language. Where two annotators were available (Czech, Slovak, Danish, and Norwegian), only sentences with unanimous agreement were kept. More details on the annotation process, including annotator counts per language and inter-annotator agreement statistics, are in Appendix A.

The resulting items fall into two subsets:

- Single: sentences valid in *only one* language but structurally similar to the other—a particularly hard case for LangID. We disregard language, for which only a few sentences remained after filtering.
- Joint: sentences valid in *both* languages of the pair.

We evaluate the 'Single' portion of the dataset with the F1 score. For the 'Joint' part of the dataset, we report the distribution of how often the model chose the respective language in language pairs vs. other languages.

3.2 Transliteration Data

Cyrillic to Latin. We transliterated Bulgarian, Macedonian, Russian, Serbian, and Ukrainian text (sourced from the Leipzig Wikipedia corpus) into Latin script using two target typographies: English (via CyrTranslit; Labrèche, 2015). We further transliterated Russian into Czech using Czech orthography (via conversion scripts from the Charles translator project; Popel et al., 2024).

Kazakh to Arabic. Kazakh has historically been written in Arabic script (before Soviet-era Cyrillisation), and standards for Arabic-script Kazakh are well established. We transliterated Cyrillic Kazakh text into Arabic using scripts from the MC^2 dataset (Zhang et al., 2024).

Homoglyph attacs. In the Latin text, letters are randomly replaced by visually identical Cyrillic or Greek characters in Czech, Slovak, and English.

Internet slang. We also include several related phenomena: (1) **diacritics removal** for Czech, Slovak, and Vietnamese; (2) **antspeak** (`s p a c e s b e t w e e n l e t t e r s`) and **leetspeak** (character substitution, e.g. `n00b`) for English, using pyleetspeak (Álvaro Huertas García, 2021).

4 Evaluation

4.1 Models

We evaluate four widely used LangID systems: **FastText** (Joulin et al., 2017), **OpenLID** (Burchell et al., 2023), **GlotLID** (Kargaran et al., 2023), and **GCLD3**. All four are multiclass classifiers that always output exactly one language label. More details on the models are in Table 5 in Appendix B.

4.2 Results on Cousin Languages

Table 3 shows the results both for the Single-language subsect (evaluation using F1 score) and the language distributions for the Joint subset. In the 'Single' subset, the scores are relatively low, given that they are high-resource languages. They lag behind the models' overall performance, indicating they are easy to confuse with the cousin languages.

The 'Joint' subset (sentences valid in both languages) reveals strong systematic bias. For

	FLORES 200	Single (F1 score)				Joint (distribution)											
		ces	spa	nob	dan	ces	slk	oth	spa	cat	oth	por	glg	oth	nob	dan	oth
FastText	.912	.84	.93	.87	.90	.38	.62	.00	.76	.15	.09	.85	.15	.00	.56	.43	.01
OpenLID	.926	.95	.96	.79	.82	.91	.09	.00	.72	.01	.27	.68	.31	.01	.67	.29	.04
GlotLID	.939	.95	.94	.88	.90	.83	.17	.00	.81	.12	.07	.97	.02	.01	.57	.43	.00
GCLD3	.352	.71	.84	.76	.83	.71	.21	.08	.55	.34	.11	.50	.47	.03	.44	.50	.06

Table 3: Language identification results for cousin language pairs under single and joint prediction. *oth* = fraction assigned to neither language in the pair. The low score of GCLD3 on the FLORES dataset is due to the smaller language coverage of the system.

	Transliteration to						Orthography noise							
	eng				ces	arb	Diacritics			Homoglyphs			Internet slang	
	bul	mkd	rus	srp	rus	kaz	ces	slk	vie	eng	ces	slk	Ant	Leet
FastText	.99⟩ .00	.99⟩ .00	.99⟩ .00	.99⟩ .00	.99⟩ .00	.99⟩ .00	1⟩ .89	1⟩ .92	1⟩ .90	.99⟩ .92	1⟩ .98	1⟩ .97	.99⟩ .12	.99⟩ .98
OpenLID	.99⟩ .00	.99⟩ .00	.95⟩ .00	.99⟩ .01	.95⟩ .00	.99⟩ .00	.99⟩ .81	.99⟩ .88	.99⟩ .94	.99⟩ .54	.95⟩ .96	.99⟩ .94	.95⟩ .00	.95⟩ .91
GlotLID	1⟩ .00	1⟩ .00	1⟩ .00	1⟩ .00	1⟩ .00	1⟩ .00	1⟩ .85	1⟩ .91	1⟩ .89	.99⟩ .77	1⟩ .97	1⟩ .97	.99⟩ .00	.99⟩ .92
GCLD3	.97⟩ .68	.97⟩ .00	.98⟩ .76	.97⟩ .00	.98⟩ .48	.99⟩ .00	.97⟩ .62	.97⟩ .74	.98⟩ .13	.93⟩ .35	.97⟩ .80	.97⟩ .72	.93⟩ .00	.93⟩ .72

Table 4: Language identification results (F1 score) for transliteration and orthography noise part of the dataset. Each cell contains the F1 score before applying the orthography noise in a smaller font, followed by ⟩ sign and the F1 score on the orthography modification of the dataset.

Czech/Slovak, OpenLID and GlotLID assign over 90% of jointly valid sentences to Czech, effectively ignoring Slovak. Surprisingly, FastText prefers lower-resourced Slovak. A similar pattern holds for Portuguese/Galician and Spanish/Catalan. GCLD3 shows less extreme bias but achieves lower overall F1, suggesting it trades accuracy for a more balanced, if still imprecise, distribution.

4.3 Results on Orthography Issues

Table 2 presents the results on the orography subset of the dataset. The F1 scores of all systems before the orthography modification are almost perfect, showing that the work with examples that are otherwise easy for LangID systems.

Transliteration is catastrophic for nearly all systems. FastText, OpenLID, and GlotLID score 0 across all Cyrillic-to-Latin conditions. They fail to recognize the language at all once the expected script is removed. GCLD3 shows partial robustness for Russian and Bulgarian transliteration into English (F1 scores of 0.76 and 0.68), though it, too, fails on Macedonian and Serbian, and, while not entirely, on Czech-orthography Russian. All systems score 0 on Arabic-script Kazakh.

In contrast, diacritics removal and homoglyph substitution have a more moderate, but still strong impact. Most systems retain F1 above 0.80 for diacritics-stripped Czech, Slovak, and Vietnamese, and above 0.9 for Czech and Slovak homoglyph attacks. OpenLID is a partial exception, dropping to 0.54 on English homoglyphs.

Antspeak proves most disruptive: FastText scores 0.12, while all other systems score 0, suggesting that inter-character spacing breaks n-gram statistics entirely. Leetspeak is handled more gracefully, with all systems achieving F1 above 0.70.

5 Conclusions

We presented CHALIS, a challenge benchmark targeting two systematic failure modes in language identification: cousin-language discrimination and non-standard orthography. Evaluating four widely used LangID systems, we found a consistent asymmetry: models reliably identify higher-resource languages while largely failing on lower-resource counterparts, with several models assigning nearly all ambiguous sentences to the dominant language in a pair. Transliteration proves catastrophic across the board, while antspeak similarly defeats current approaches, and diacritics removal and homoglyph substitution are handled more gracefully. These findings suggest that current LangID systems overfit to surface-level script and frequency cues. We hope CHALIS will motivate the development of stronger identification systems for lower-resource languages whose speakers deserve reliable NLP tooling.

Limitations

CHALIS covers only four cousin language pairs, all Indo-European languages from Europe. While these pairs illustrate systematic LangID failure modes, the benchmark does not capture the full diversity of closely related language families, such as Arabic dialects, Turkic languages, or South and Southeast Asian language groups, where resource disparities and mutual intelligibility raise similar concerns.

CHALIS is diagnostic by design and frames LangID as multiclass classification, so the 'Joint' subset necessarily requires models to commit to a single label even for sentences that are valid in both cousin languages, and there is no single correct answer for such cases. A *multilabel* formulation, i.e., asking independently whether a sentence could be each candidate language, would allow reasonable scores for both languages and better suit the 'Joint' subset. We see this as a promising direction, but it is out of scope for the present, diagnostic benchmark.

We evaluate CHALIS at the sentence level, which makes for a challenging and, we believe, informative setup, but is stricter than most realistic data-filtering pipelines, where longer context generally makes classification easier. Still, sentence-level distinctions matter in their own right in some settings, e.g., bilingual documents in territories where cousin languages are co-official.

The human annotation phase, which validates the 'Single' and 'Joint' subsets, was conducted by a small number of annotators. The orthography subset was constructed fully automatically and does not reflect the natural distribution of transliteration or slang usage found in real web-crawled data. Transliteration standards vary across communities and over time, and our choices (e.g., CyrTranslit for Cyrillic-to-English, Czech orthography for Russian) represent only a subset of conventions in actual use. Similarly, antspeak and leetspeak were generated programmatically rather than sampled from genuine user-generated content.

Finally, CHALIS is an evaluation-only benchmark. It does not include a training signal, so it cannot be used directly to improve LangID systems; it can only be used to diagnose them. Addressing the failure modes we identify will require targeted data collection and model development efforts beyond the scope of this work.

Acknowledgments

We thank Jindra Helcl for interesting discussions on language identification and feedback on early drafts of this paper.

This research was supported by the Charles University project PRIMUS/23/SCI/023 and project CZ.02.01.01/00/23_020/0008518 of the Ministry of Education, Youth and Sports of the Czech Republic. Computational resources were provided by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

We used GitHub Copilot when writing code. We used Grammarly for grammar checking, Claude for proofreading, and for text shortening.

References

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, Toronto, Canada. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 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 (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Isaac Caswell, Theresa Breiner, Daan Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. pages 6588–6608.

J. K. Chambers and Peter Trudgill. 1998. *Dialectology*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic language identification in texts: A survey. *CoRR*, abs/1804.08186.

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. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184 vol.1.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Georges Labrèche. 2015. Cyrtranslit. Version 1.1.1, MIT Licence.

Alexander Maxwell. 2018. When theory is a joke: The weinreich witticism in linguistics. *Beitrage zur Geschichte der Sprachwissenschaft*, 28:263–292.

Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *J. Comput. Sci. Coll.*, 20(3):94–101.

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*.

Martin Popel, Lucie Polakova, Michal Novák, Jindřich Helcl, Jindřich Libovický, Pavel Straňák, Tomas Krabac, Jaroslava Hlavacova, Mariia Anisimova, and Tereza Chlanova. 2024. Charles translator: A machine translation system between Ukrainian and Czech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3038–3045, Torino, Italia. ELRA and ICCL.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. MC^2: 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, Bangkok, Thailand. Association for Computational Linguistics.

Álvaro Huertas García. 2021. Pyleetspeak. Version 0.3.9, MIT Licence.

A Annotation Procedure Details

Annotators were university students recruited via internal social media. We explained to them what data we want to annotate for what purpose. The annotators were asked to decide whether a sentence is in their native language. We estimated that processing 1,200 sentences would take approximately one hour and offered the students a compensation

Model	Architecture	Model size	License
FastText (lid218e)	FastText	1.2 GB	CC BY-NC 4.0
OpenLID	FastText	1.3 GB	GPL 3.0
GlotLID (v3)	FastText	1.7 GB	Apache 2.0
GCLD3	Char-n-gram-based NN	15MB	Apache 2.0

Table 5: Overview of evaluated LangID systems. Model sizes are taken from the `.bin` file, where available. FastText-based models do not report discrete parameter counts.

of 500 CZK (3.7× the minimum wage and 2× the median wage in 2025 in Czechia).

For Czech, Slovak, Danish, and Norwegian, two annotators evaluated each language. Out of the 1,200 sentences annotated per pair, disagreements were rare, between 2 and 15 cases per language. These were resolved by presenting the disputed sentences to both annotators and asking them to reach a consensus. For Spanish, Catalan, Portuguese, and Galician, we had only one annotator per language.

B Licenses of Software and Data

Table 5 summarizes the licenses of the four evaluated LangID systems. All evaluated models are freely available for research use: FastText (lid218e) is released under CC BY-NC 4.0, restricting commercial use; OpenLID under GPL 3.0; and both GlotLID and GCLD3 under the permissive Apache License 2.0. The training corpora used to construct CHALIS are similarly open: WMT News Crawl data and the Leipzig Corpora Collection are freely available for research purposes, and Wikipedia-derived data is released under CC BY-SA 4.0. The KenLM language modeling toolkit is distributed under the LGPL 3.0 license. The GLiNER named-entity recognizer used for filtering is released under the Apache License 2.0. Transliteration tools include CyrTranslit (MIT license), the Charles Translator conversion scripts, and the MC^2 dataset scripts; pyleetspeak is also MIT-licensed. The Moses tokenizer is distributed under the LGPL 2.1 license. CHALIS itself will be released under CC BY 4.0 to maximize reusability.

C Dataset Format

CHALIS is distributed as a single JSON file structured as a dictionary with the following top-level keys: `single`, `joint`, `cyrilic_to_czech`, `cyrilic_to_english`, `antspeak`, `arabic`, `diacritics`, `homoglyphs`, and `leet`. Each key maps to a list of items, where every item is a dictionary with the following fields:

- `text`: the string to be classified;
- `label`: a list of language labels in BCP-47-style format extended with script information (e.g., `['ces_Latn', 'slk_Latn']`), following the OpenLID convention. For the `joint` subset, the list contains two labels; for all other subsets, it contains exactly one;
- `original`: the source text prior to rule-based modification. This field is present in all subsets except `single` and `joint`, where the texts were collected directly rather than derived from another text.

The CreateDebate Arguments Dataset: A Large-Scale Naturalistic Corpus for Computational Argumentation

Antonis Charalampous and Constantinos Djouvas
Department of Communication and Internet Studies
Cyprus University of Technology
ag.charalampous@edu.cut.ac.cy, constantinos.djouvas@cut.ac.cy

Abstract

We introduce the CreateDebate Arguments Dataset, a large-scale, unfiltered collection of 383,220 arguments from 7,435 of the most active for/against debates on the CreateDebate online debate platform, spanning 16 years of activity. Unlike existing computational argumentation resources, the dataset combines large scale with hierarchical debate-tree structure, author-assigned stance labels, community vote scores, and relational tags that explicitly mark when one argument disputes another. Because the full dataset is unfiltered, we release it under a gated access agreement and additionally release an openly available, manually curated subset of 84,872 arguments from 578 debates. To demonstrate the dataset's utility, we define two counter-argument retrieval benchmarks on the curated subset: a *targeted* setting that uses the dispute tag as author-verified ground truth, and a *general* setting in which any opposing-side argument is a valid counter. Evaluating a BM25 baseline, a frozen bi-encoder, and a fine-tuned bi-encoder, we find consistent but bounded gains from fine-tuning on disputed pairs, a negative effect of debate-level context on within-debate retrieval, and limited cross-task transfer between the targeted and general settings. A qualitative analysis surfaces failure modes specific to naturally occurring debate. The dataset and models are released to the community.

1 Introduction

The rise of online debate platforms, social media, and digital news has made argumentation an increasingly central phenomenon in public discourse. As a result, the automatic analysis of argumentative content, broadly referred to as *computational argumentation*, has attracted considerable attention in the natural language processing community (Stab and Gurevych, 2014; Lippi and Torroni, 2016). Research in this area spans a wide range of tasks, from identifying argumentative components within documents to assessing argument quality and detecting stance.

Two broad paradigms have emerged. *Discourse-level argument mining* focuses on identifying the internal argumentative structure of a document and recovering the inferential relations between its parts (Stab and Gurevych, 2014). *Information-seeking argument mining* takes a wider view: rather than analyzing a single document in depth, it aims to identify self-contained argumentative statements relevant to a given topic or query from large, heterogeneous text collections (Trautmann et al., 2020). In an information landscape saturated with one-sided content, there is a growing need for tools that surface diverse argumentative perspectives on controversial topics on demand (Levy et al., 2018; Ajjour et al., 2019). Existing resources each make different trade-offs between scale, structural richness, naturalness, and metadata granularity. We position our dataset as an additional point in this trade-off space: a large-scale collection of naturally occurring debate threads with rich metadata supporting a wide range of computational argumentation tasks.

We introduce the **CreateDebate Arguments Dataset**, a large-scale collection of **383,220 arguments** across **7,435 debates** collected from CreateDebate, an online platform where users post arguments, reply to one another, and vote on argument quality. Arguments are organized in hierarchical debate trees and carry author-assigned stance labels, community vote scores, and relational tags that explicitly mark when one argument disputes another. This combination of scale, structural richness, and naturally occurring metadata makes the dataset suitable for stance detection, argument quality assessment, discourse structure analysis, and counter-argument retrieval. The collection targets the most active debates of a single format on the platform rather than its entire content (Section 3).

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 8–21
October 29, 2026 ©2026 Association for Computational Linguistics

Active since 2008, CreateDebate combined an open debate community with a classroom mode in which instructors could set up private debate communities for their students, and it was promoted as a venue for practising persuasive writing; particularly in its early years, part of its user base therefore plausibly consisted of students and educators. Its debates have supported research on stance and reason classification (Hasan and Ng, 2014), dialogic argumentation (Abbott et al., 2016), and explainable stance detection (Saha et al., 2024). Because the platform is no longer accessible, a large collection of its debates also has documentary value: spanning 16 years, it makes it possible to study how participation, topics, and argumentative register evolved on a single platform, including shifts such as the growth of trolling and low-effort contributions.

Because the dataset is unfiltered, it necessarily contains debates on extremely sensitive and potentially disturbing topics, alongside the large majority of debates that are not. We therefore release the full dataset under a gated access agreement restricting its use to research purposes, and additionally release a smaller, manually curated subset of **84,872 arguments** across **578 debates**, selected for topical coherence and quality. Details of the access procedure and curation methodology appear in Section 3.

To demonstrate the dataset's utility, we focus on *counter-argument retrieval*: given an argument advancing a position, retrieve from a corpus the arguments that most effectively oppose it (Wachsmuth et al., 2018; Orbach et al., 2020; Jo et al., 2021). We conduct this demonstration on the curated subset, since the task does not benefit methodologically from the most sensitive content. The dataset's *disputed* tag provides a naturally occurring, author-verified supervision signal directly suited to this task. We introduce two retrieval benchmarks: a *targeted* setting that uses the disputed tag as strict, author-verified ground truth, and a *general* setting in which any opposing-side argument is a valid counter. We evaluate three systems: a BM25 lexical baseline, a frozen pre-trained bi-encoder, and a fine-tuned bi-encoder trained with contrastive learning, and report findings on the role of debate context, hard-negative mining, and the relationship between the two settings.

Our contributions are:

- We release a large, unfiltered collection of 383,220 arguments from 7,435 debates under a gated access agreement, with stance labels, relational tags, community vote scores, and hierarchical tree structure[1].
- We release a manually curated subset of 84,872 arguments from 578 debates, suitable for open downstream use.[2]
- We demonstrate the dataset on two counter-argument retrieval tasks and release the resulting models[3].

2 Related Work

2.1 Computational Argumentation Datasets

The development of computational argumentation has been closely tied to the availability of dedicated corpora. The Internet Argument Corpus 2.0 (Abbott et al., 2016) aggregates posts from three debate platforms (4forums.com, Convinceme.net, and CreateDebate.com) with annotations for stance, agreement, and dialogue structure. Its CreateDebate subset comprises 2,958 posts and lacks the fine-grained relational metadata, such as explicit dispute relations and hierarchical tree structure, that characterizes platform-native debate threads. CreateDebate has also been a recurring source for stance research. Hasan and Ng (2014) annotate ideological debate posts from four domains with the reasons behind each author's stance and jointly model stance and reason classification, while Saha et al. (2024) crawl CreateDebate, together with Room For Debate, to build claim–article benchmarks for explainable stance detection (about 66,000 samples across the two sources). These resources are organized around individual posts or claim–document pairs for stance classification, rather than around the reply structure and relational tags between arguments. Our dataset draws substantially more content from the same platform while preserving its full structural richness.

The args.me corpus (Ajjour et al., 2019) harvests 387,606 arguments from four debate portals using distant supervision over explicit pro/con structure, but its arguments are drawn from pre-structured schemas with predefined stances, limiting naturalness and relational diversity. The accompanying

[1] https://huggingface.co/datasets/azza1625/createdebate-arguments-full

[2] https://huggingface.co/datasets/azza1625/createdebate-arguments

[3] https://huggingface.co/azza1625/counter-argument-retrieval

args.me search engine (Wachsmuth et al., 2017b) and TARGER (Chernodub et al., 2019) demonstrate the value of such corpora for argument retrieval over web-scale collections. Our dataset complements these efforts with a resource grounded in naturally occurring debate interactions and richer relational structure.

IBM's Project Debater program produced several influential resources (Levy et al., 2018; Slonim et al., 2021). Most relevant is the dataset of Orbach et al. (2020), comprising over 3,600 professionally recorded debate speeches annotated for counter-speech relations; it is restricted to formal speeches by expert debaters. The ArguAna Counterargs corpus (Wachsmuth et al., 2018), derived from idebate.org, provides 6,753 argument–counterargument pairs across 1,069 debates and is a key counter-argument retrieval benchmark, but its scale and reliance on a single structured portal constrain stylistic and topical diversity. Counter-argument generation work has frequently used the ChangeMyView dataset (Jo et al., 2021; Alshomary et al., 2021; Alshomary and Wachsmuth, 2023), which preserves threaded structure but provides no explicit debate sides or dispute tags. A related line studies argument quality (Wachsmuth et al., 2017a; Habernal and Gurevych, 2016; Gretz et al., 2020), relevant to ours given the community vote scores in our dataset.

2.2 Argument Retrieval and Generation

Work beyond classification follows two complementary directions: retrieval of relevant arguments and generation of novel argumentative content, both situated within the information-seeking paradigm (Trautmann et al., 2020). Systems such as ArgumenText (Stab et al., 2018; Si et al., 2023) show that identifying argumentative content from web-scale corpora is feasible and that topic information improves retrieval. On the retrieval side, Wachsmuth et al. (2018) model the similarity and dissimilarity of argument pairs to find the best counterargument, defining eight retrieval tasks of varying complexity over ArguAna. Orbach et al. (2020) frame counter-argumentation as detection (identifying which opposing speech most directly counters a given one), and find that expert humans substantially outperform automatic methods. Jo et al. (2021) retrieve counter-evidence to specific statements using a knowledge-enhanced NLI model, while Le et al. (2018) build retrieval- and generation-based argumentative dialogue agents on the Internet Argument Corpus. On the generation side, CANDELA (Hua et al., 2019) combines retrieval with a two-step neural generator; Alshomary et al. (2021) attack weak premises to undermine arguments; and Alshomary and Wachsmuth (2023) jointly model conclusion and counter-argument generation with a stance-based ranking component.

Prior datasets occupy different points in a trade-off space defined by scale, structural richness, naturalness, and metadata granularity. The CreateDebate Arguments Dataset complements these resources by offering substantial scale, naturally occurring hierarchical conversational structure, and author-tagged relational metadata between argument pairs within a single platform; to our knowledge, it is the largest public CreateDebate corpus that preserves the full reply tree together with the author-assigned relational tags.

3 The CreateDebate Arguments Dataset

We begin by introducing the CreateDebate platform and the two-phase collection procedure, then describe the curated subset and a descriptive analysis of the dataset's structural characteristics.

3.1 Platform and Data Collection

The CreateDebate Platform. CreateDebate lets users initiate debates on an unrestricted range of issues. A debate begins with a user-submitted proposition, optionally with a contextual description, and takes one of three forms: open-ended, for/against, or multiple-sides. Discussions unfold as hierarchical trees: arguments receive replies, forming multi-level conversational structures of arbitrary depth. Replies are categorized into three interaction types, *Support*, *Dispute*, and *Clarify*, providing explicit indicators of argumentative relations between posts. The platform also supports up/downvoting of individual arguments, with scores displayed alongside the text, and records debate-level metadata such as total and side-specific scores. These properties make CreateDebate a rich source for argumentative datasets (Levy et al., 2014; Ajjour et al., 2019; Habernal and Gurevych, 2016).

Phase One: Debate Discovery. Collection used automated crawlers, with Selenium driving the dynamic browser interactions required to navigate the platform. The first phase identified the set of debates to extract. We set the topics filter to include all categories, surfacing debates from politics and religion to sports, technology, and entertainment,

and sorted debates in descending order by argument count. We restricted collection to the *for/against* format, which provides a more structured argumentative setting than open-ended or multi-sided debates, and included both open and closed debates. The crawler iterated through the first 100 listing pages, yielding **7,435 unique debates** after deduplication and removal of inaccessible debates. The collection is therefore not an exhaustive crawl of CreateDebate: it covers the most active for/against debates reachable through these listing pages, and excludes open-ended and multi-sided debates as well as less active ones. Because the platform is no longer accessible, the total number of debates it hosted cannot be established, so we do not report a coverage ratio. Throughout the paper, *full dataset* refers to this unfiltered collection, as opposed to the curated subset. Basic metadata (URL, declared winning position when available, and platform-assigned topic) was recorded at this stage.

Phase Two: Argument Extraction. The second phase visited every debate URL to retrieve arguments and metadata. For each argument node we extracted the text, timestamp, user-declared stance, community vote score, depth in the tree, and the relational tag of any reply (*Support*, *Dispute*, or *Clarify*). At the debate level we recorded the title, full description, and aggregate statistics. Extraction yielded **383,220 arguments** across the 7,435 debates. The structure of the resulting debate- and argument-level objects is illustrated in Appendix A (Tables 6 and 7).

3.2 The Curated Subset

Alongside the full dataset we release a manually curated subset of 84,872 arguments across 578 debates, intended for open downstream use without access restrictions.

Selection Methodology. Because CreateDebate is user-generated and minimally moderated, the full dataset includes debates on inappropriate or sensitive topics and near-duplicate debates. The curated subset was produced by manual review at the debate level, carried out by the two authors by examining debate titles. The review had two main goals: filtering out NSFW debates, and making the retained debates as topically varied as possible. In practice, a debate was excluded if (i) its title indicated sexual, graphic, or otherwise NSFW content; (ii) it posed essentially the same proposition as a debate already selected, which removed near-duplicates and served the variety goal; or (iii) its title did not pose a coherent, debatable proposition. Near-duplicates were identified by manually comparing titles rather than by an automatic similarity threshold. During the same pass, the reviewers assigned topic keywords to each retained debate to organize the subset and support topic-based filtering. Rather than imposing a post-level definition of a valid argument, the subset captures online debate as it actually unfolds, making it more representative of real-world conditions while introducing challenges absent from more controlled corpora.

Table 1: Corpus-level statistics of the full dataset and the curated subset. †Mean (median). ‡Mean (maximum).

Statistic	**Full**	**Curated**
Debates	7,435	578
Arguments	383,220	84,872
Unique users	26,350	10,869
Vocabulary size	459,726	164,673
Arguments per debate†	51.5 (35.0)	146.8 (117.5)
Argument length (words)†	77.3 (38.0)	92.6 (45.0)
Tree depth‡	4.27 (109)	5.28 (109)
Avg. branching factor	1.28	1.32
Replies (% of arguments)	72.7	73.3
Disputed (% of replies)	70.0	77.5
Balanced debates (%)	56.3	69.8
Date range	Feb 2008 – Mar 2024	

Additional Annotations. Besides the topic keywords, we computed two structural fields for each curated debate. The *depth* field gives each argument's position in the tree (0 for top-level), and the *context* field stores the parent argument's full text for replies. These fields allow the full tree structure of any debate to be reconstructed from the flat representation, and are also present in the full release.

3.3 Dataset Statistics and Analysis

Table 1 summarizes both releases; we first describe the full dataset. It spans 16 years of activity (February 2008–March 2024), comprising 383,220 arguments from 26,350 users across 7,435 debates, with an average of 51.5 arguments per debate (median 35.0) and a vocabulary exceeding 459,000 tokens. Notably, 72.7% of all arguments are replies rather than top-level posts, underscoring the strongly conversational nature of the corpus.

Figure 1 characterizes the full dataset along four dimensions, and Figure 2 (Appendix B) adds two more; panels are labelled (a)–(f) consistently across

both figures.

(a) Argument length. The distribution is strongly right-skewed (mean 77.3 words, median 38.0): 79.8% of arguments contain fewer than 100 words and 92.2% fewer than 200, confirming that contributions are predominantly concise, characteristic of web-based discourse rather than formal debate, and fall comfortably within standard transformer input limits.

(b) Argument depth. The first three depth levels account for 55.1% of the corpus, but the dataset nonetheless reaches a mean depth of 4.27 (90% within depth 11, maximum 109), and the average branching factor of 1.28 indicates debates structured around elongated chains rather than broad, shallow trees.

(c) Arguments per debate. The number of arguments per debate is right-skewed (mean 51.5, median 35.0); 31.6% of debates exceed 50 arguments and 10.1% exceed 100.

(d) Relational tags by depth. The Disputed tag dominates throughout, accounting for 70.0% of all replies overall and rising from 66.1% at depth 1 to 71–74% from depth 3 onwards, indicating that interactions become more confrontational beyond the first reply levels. The Disputed tag reflects the reply author's intent to challenge rather than a guarantee of argumentative substance: in its natural form, some disputed replies are brief reactions or off-topic remarks rather than substantive counters. This property directly shapes the targeted retrieval benchmark in Section 4, where disputed pairs serve as author-verified ground truth.

(e) Chain length and (f) side balance. Chain length, the number of exchanges from root to a given argument (i.e., depth + 1), is near-uniformly distributed across top-level (27.3%), short (27.8%), medium (20.7%), and deep (24.2%) bands, with over 72% of arguments embedded in exchanges of length two or more (mean 5.27, maximum 110), confirming that the dataset is genuinely dialogic rather than a collection of independent statements. Side balance, the fraction of arguments from the more active side, has mean 0.609 and median 0.588: 56.3% of debates are balanced (ratio 0.5–0.6), 29.2% moderately imbalanced, and 14.5% strongly imbalanced, so most debates are reasonably contested.

Table 2: Computational argumentation tasks supported by the dataset's fields. † = task demonstrated in this paper. Multiple fields are separated by semicolons.

Task	Field(s)	Description
Stance detection	Side; Body	Predict argument's side
Argument quality	Points	Rank by vote score
Relation classif.	Tag	Classify reply relation
Counter-arg. retr.†	Tag (Disputed); Side	Retrieve opposing counter
Dialogical mining	context; depth	Model multi-turn exchange
Outcome prediction	Total Score; Side Points	Predict winning side

Comparison with the Curated Subset. Table 1 also contrasts the full dataset and the curated subset. The two are similar in shape but differ in scale and a few properties: the curated subset has larger debates (146.8 vs. 51.5 arguments per debate), a higher disputational rate (77.5% vs. 70.0% of all replies), and tighter side balance (69.8% vs. 56.3% balanced debates), reflecting the selection bias toward larger, more substantive, contested debates. The share of replies and the branching factor are nearly identical across the two releases (73.3% vs. 72.7%; 1.32 vs. 1.28), while curated debates are somewhat deeper (mean depth 5.28 vs. 4.27) and their arguments slightly longer (median 45 vs. 38 words).

3.4 Supported Tasks

The combination of fields in Table 2 supports a range of computational argumentation tasks beyond the counter-argument retrieval benchmarks demonstrated in Section 4. The tasks selected serve to highlight how a few of the dataset's distinguishing fields such as argument and user vote scores, relational dispute tags, and dialogical structure, each correspond to a different computational argumentation task. Many other tasks (e.g. argument retrieval, evidence detection, conclusion generation) could also be formulated over this data.

4 Counter-Argument Retrieval

We present counter-argument retrieval as a demonstration task, conducted on the curated subset. The curated subset suffices because the task does not benefit methodologically from the most sensitive content, allowing us to release trained models and evaluation data without restrictions, and because it is enriched for larger, more substantive debates

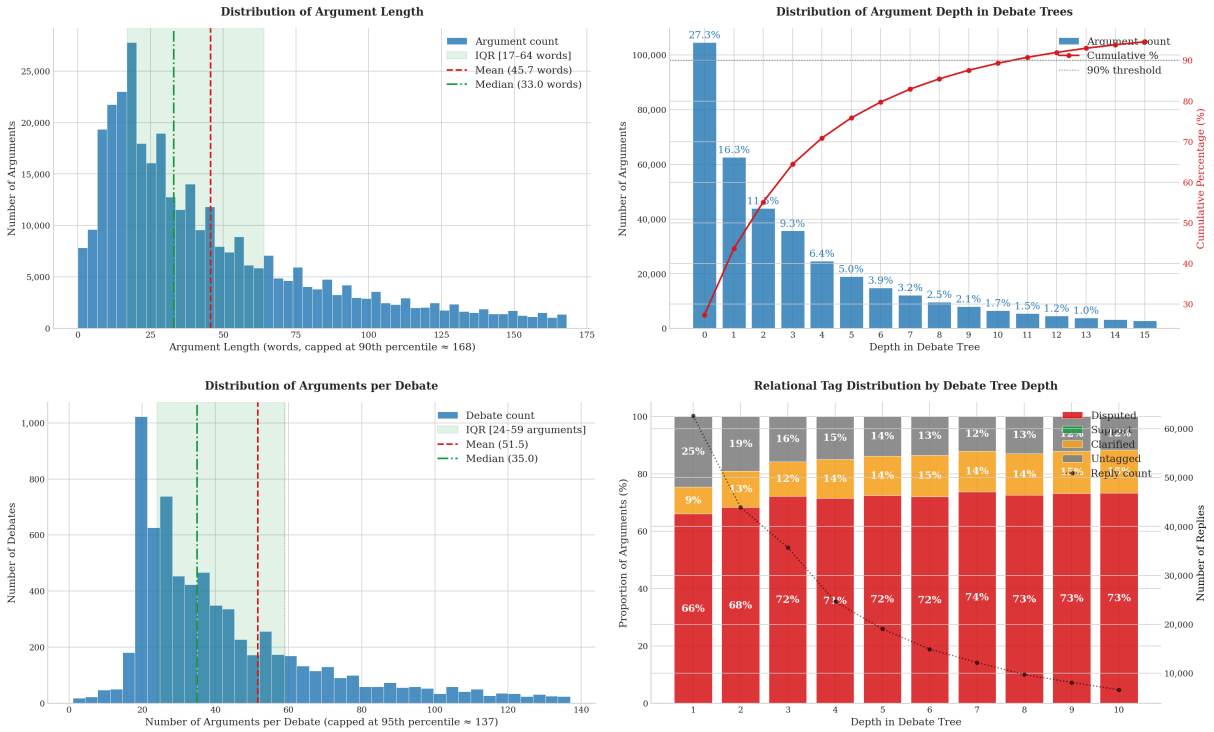


Figure 1: Structural statistics of the full CreateDebate Arguments Dataset: (a) argument length, (b) argument depth in debate trees, (c) arguments per debate, and (d) relational tag proportions by depth. Panels (e) chain length and (f) side balance appear in Figure 2 (Appendix B). In (a) and (c), the x-axis is capped for display, and the mean and median lines in (a) are computed over the displayed range.

that yield candidate pools of sufficient size and topical coherence. Among the tasks in Table 2, we select counter-argument retrieval because it most directly exercises the dataset's most distinctive field, the author-assigned Disputed tag, which side-label-only corpora cannot provide. The experiments address three questions: (RQ1) does fine-tuning on disputed pairs improve targeted retrieval over lexical and off-the-shelf semantic baselines; (RQ2) does debate-level context help retrieval within a debate; and (RQ3) do targeted and general retrieval share a transferable signal?

4.1 Task Formulation

Given a query argument q advancing a position, counter-argument retrieval retrieves from a candidate pool $\mathcal{C}$ the arguments that most effectively oppose it. The dataset supports two instantiations differing in how ground truth is defined.

Scenario 1 – Targeted Retrieval. This setting uses the *disputed* tag for author-verified ground truth. For each disputed reply we construct an instance whose query is the parent argument, whose positive is the disputed reply itself (which targets the opposing side in 95.1% of cases), and whose negatives are all other opposing-side arguments within the same debate. The model must identify the specific argument authored as a direct rebuttal to the query from among all opposing-side arguments. After filtering same-side disputed pairs (4.9%), this yields *44,569* query–positive pairs.

Scenario 2 – General Retrieval. This setting treats any opposing-side argument as a valid counter. For each argument, the query is any argument from side A, the positives are all arguments from side B, and the negatives are all other side-A arguments in the same debate. It tests opposition detection at the debate level rather than the claim level. For debates with more than two sides we keep the two dominant sides (dropping 2,466 fringe arguments, 2.9%). After a per-debate cap of 500 pairs, this yields *80,885* query instances.

Data Splits. Both scenarios share a debate-level 75/10/15 train/validation/test split, so all arguments from a debate fall in a single partition, preventing topic-specific leakage, analogous to the train–test overlap documented for open-domain QA by Lewis et al. (2021). Not all debates contribute to both scenarios. Per-split instance counts appear in Appendix C. The average negative pool of roughly

94–120 arguments per query in Scenario 1 makes the task non-trivial, as all candidates are drawn from the same debate context.

4.2 Models

We evaluate three systems of increasing sophistication.

BM25. A lexical baseline (Robertson and Zaragoza, 2009) requiring no training and operating on surface term overlap, a long-standing strong IR baseline (Thakur et al., 2021). We use the `rank_bm25` implementation with default parameters and whitespace tokenization.

Frozen Bi-Encoder. A pre-trained sentence encoder without fine-tuning (Reimers and Gurevych, 2019). Query and candidates are encoded independently and ranked by cosine similarity. We use `all-mpnet-base-v2`, establishing how far off-the-shelf semantic representations suffice without domain adaptation.

Fine-Tuned Bi-Encoder. Our primary model fine-tunes the same backbone with MultipleNegativesRankingLoss (Henderson et al., 2017), a contrastive objective treating other in-batch positives as implicit negatives. We also train a variant with one explicit hard negative sampled from the opposing pool. Training uses batch size 64, learning rate 2×10^{-5}, and 3 epochs with 10% linear warmup; the best checkpoint is selected by MRR@10 on validation. We train separate models per scenario (four variants total) and evaluate each on both test sets, using the `sentence-transformers` library (Reimers and Gurevych, 2019).

4.3 Evaluation Protocol

For **Scenario 1**, each query has a single ground-truth positive, so we report MRR (primary) and Recall@k for $k \in \{1, 5, 10\}$. For **Scenario 2**, each query has multiple valid positives, so we report Recall@k for $k \in \{10, 20\}$ at larger k, since the positive pool (mean 83.3) makes low-k recall structurally bounded. Candidate corpora are encoded once per model and scored by dot product between normalized embeddings. For each model we evaluate two input configurations: argument text only, and text prepended with the debate title and description to test whether debate-level context aids retrieval.

Table 3: Scenario 1 (Targeted Retrieval). Best per column in **bold**. Fine-tuned models use argument text only.

Model	Train	MRR	R@1	R@5	R@10
Baselines					
BM25	w/ ctx	0.277	0.182	0.353	0.439
BM25	no ctx	0.362	0.249	0.471	0.562
Frozen	w/ ctx	0.170	0.084	0.224	0.324
Frozen	no ctx	0.381	0.253	0.517	0.631
Fine-tuned on Scenario 1					
Bi-enc.	In-batch	**0.444**	**0.301**	**0.605**	**0.711**
Bi-enc.	Hard neg.	0.444	0.300	0.603	0.710
Fine-tuned on Scenario 2 (transfer)					
Bi-enc.	In-batch	0.365	0.242	0.490	0.594
Bi-enc.	Hard neg.	0.363	0.240	0.491	0.596

Table 4: Scenario 2 (General Retrieval). All systems lie within 0.002 of each other, so no best system is marked.

Model	Train	R@10	R@20
Baselines			
BM25	—	0.079	0.159
Frozen	—	0.078	0.158
Fine-tuned on S1 (transfer)			
Bi-enc.	In-batch	0.077	0.157
Bi-enc.	Hard neg.	0.077	0.157
Fine-tuned on S2			
Bi-enc.	In-batch	0.077	0.157
Bi-enc.	Hard neg.	0.077	0.157

4.4 Results and Discussion

Tables 3 and 4 report results for all variants.

Effect of Debate Context. Adding debate context consistently degrades Scenario 1 performance: BM25 drops from 0.362 to 0.277 MRR and the frozen encoder from 0.381 to 0.170. Since all candidates come from the same debate, prepending the title introduces shared vocabulary that inflates similarity uniformly, reducing the discriminative signal. At inference time, debate context should therefore not be included when candidates are drawn from the same debate.

Scenario 1 Results. Fine-tuning on disputed pairs improves over the frozen encoder by 6.3 MRR and 8.0 R@10 points (MRR 0.444, R@10 0.711). These gains are more modest than is typical in bi-encoder fine-tuning (Thakur et al., 2021), likely for two reasons: all negatives are drawn from the same debate and are thus already topically similar, so the frozen encoder captures much of the needed discrimination; and the training data mixes substantive rebuttals with non-substantive exchanges, since the disputed tag captures intent rather than

quality. Hard-negative training yields no measurable improvement, as a batch size of 64 already provides 63 diverse implicit negatives per query.

Cross-Scenario Transfer. Training on Scenario 2 and evaluating on Scenario 1 falls below the frozen encoder baseline (MRR 0.363–0.365 vs. 0.381): optimizing for general side-opposition is mildly detrimental to targeted rebuttal retrieval. The two objectives partially compete; distinguishing opposing- from same-side arguments is a different task from identifying a specific rebuttal. Transfer in the opposite direction also provides no benefit.

Scenario 2 Results. Results are uniform across all models (differences below 0.002), with BM25 matching or marginally outperforming all neural models. This points to a fundamental difficulty with the general setting: with a mean positive pool of 83.3 and all candidates sharing the debate context, the textual signal to distinguish valid counters from same-side arguments appears insufficient for any evaluated approach. We regard these results as a lower bound on the difficulty of general counter-argument retrieval on this dataset.

Summary. Fine-tuning on author-verified disputed pairs yields consistent but bounded improvement for targeted retrieval; general retrieval as formulated is difficult for all models, with BM25 remaining competitive; and the two settings are distinct tasks with non-transferable signals, underscoring the value of the disputed tag as a supervision source enabling a qualitatively different setting than side-label-only corpora.

4.5 Qualitative Analysis

Aggregate metrics do not reveal the nature of individual successes and failures. A recurring theme is a tension inherent to the dataset: CreateDebate users produce the full spectrum of argumentative behavior, from substantive rebuttals to brief dismissals and off-topic responses. This naturalistic breadth distinguishes the dataset but also challenges retrieval models optimized for semantic similarity.

Outcome Distribution. Table 5 bins test queries by ground-truth rank. The model retrieves the correct counter at rank 1 for 29.8% of queries and within the top 5 for 60.4%; a further 20.7% are partial failures (rank 6–20) and 18.8% complete failures (rank >20, mean 48.8). The median ground-truth rank is 3, so near misses are more common than complete failures. The similarity statistics reveal the core failure mode: as retrieval quality degrades (Bins B–D), query–ground-truth semantic similarity drops sharply (to 0.298 in Bin D), yet similarity between query and the model's top-1 choice stays high (0.655–0.672). The model confidently retrieves something semantically close to the query, just not the author-verified ground truth. Failures thus arise from a mismatch between semantic similarity and the pragmatic logic of the actual rebuttal.

Table 5: Test queries by ground-truth rank, with mean lexical (Jaccard) and semantic (cosine) similarity between query and the ground truth (gt) and top-1 retrieved argument (t1).

			Lexical		Semantic	
Bin	**N**	**%**	**gt**	**t1**	**gt**	**t1**
A: Rank 1	1,157	29.8	0.251	0.251	0.672	0.672
B: Rank 2–5	1,187	30.6	0.176	0.224	0.550	0.655
C: Rank 6–20	803	20.7	0.118	0.180	0.416	0.618
D: Rank >20	731	18.8	0.092	0.179	0.298	0.622

Success and Failure Patterns. The model often succeeds despite minimal lexical overlap: across representative cases (Appendix D, Table 10) it identifies counters addressing the argumentative thrust of the query (lexical overlap 0.063–0.152, semantic similarity 0.776–0.835), confirming it operates on semantic rather than surface patterns. Failures cluster into two types. *Pragmatic and contextual rebuttals* (Table 11) engage with the framing or a specific embedded claim rather than the overall topic; these carry low or indistinguishable semantic similarity, and the single ground-truth assumption underestimates capability where multiple valid counters exist, so Scenario 1 scores are lower bounds. *Non-substantive exchanges* (Table 12) arise because any reply marked Disputed becomes ground truth regardless of substance; failing to retrieve a brief conversational reaction is arguably correct from a retrieval standpoint. These cases follow from how the data was collected, and the conversational noise in training partly explains the modest fine-tuning gains.

Error Patterns by Depth and Length. Performance at depth 0 is lower (MRR 0.334) than at deeper levels (0.416–0.497; Appendix D, Figure 3). Top-level arguments are broad opening claims plausibly countered by many opposing argu-

ments, whereas deeper exchanges are more specific and contextually grounded, narrowing the space of plausible counters. A modest positive trend also holds for query length (MRR 0.436 for 1–20 words, rising to 0.498 for 201–500 words): longer queries provide more signal, while short rhetorical remarks tend to have less substantive disputed replies.

5 Conclusion and Future Work

We introduced the CreateDebate Arguments Dataset, a large-scale, unfiltered collection of 383,220 arguments from 7,435 debates spanning 16 years, distinguished from existing resources by its combination of scale, hierarchical tree structure, author-assigned stance labels, community vote scores, and relational dispute tags. The full dataset is released under gated access reflecting its unfiltered character, accompanied by an openly available curated subset of 84,872 arguments from 578 debates. To demonstrate its utility we introduced two counter-argument retrieval benchmarks and evaluated three systems. Fine-tuning on disputed pairs yields consistent but bounded gains on targeted retrieval; adding debate context degrades within-debate retrieval; and targeted and general retrieval are distinct tasks with non-transferable signals. A qualitative analysis identified three failure modes specific to naturally occurring debate: pragmatic and meta-argumentative rebuttals below the resolution of semantic similarity, ambiguity in the single ground-truth assumption, and non-substantive exchanges carrying the disputed tag.

Several directions follow from these findings. Retrieval is inherently limited to indexed arguments, so generation-based methods (Hua et al., 2019; Alshomary et al., 2021; Alshomary and Wachsmuth, 2023) would allow novel counters. A stance classifier as a post-retrieval verification step (Alshomary and Wachsmuth, 2023) could ensure retrieved arguments genuinely oppose the query. Extending retrieval to cross-debate candidates, supported by the topic keywords, would better reflect real-world conditions. Automatic argument-level quality filtering could separate substantive arguments from conversational noise, and comparing filtered and unfiltered training would quantify the effect of naturalistic noise. Finally, relevance-graded annotation beyond single ground truth would enable fairer evaluation where multiple valid counters exist.

Limitations

The full dataset is not an exhaustive crawl of CreateDebate: it covers the most active for/against debates, so its statistics characterize that population rather than the platform as a whole. The curated subset reflects the judgment of two reviewers working from debate titles; we did not measure inter-annotator agreement, and the subset differs in distribution from the full dataset (Table 1). The counter-argument retrieval demonstration is conducted only on the curated subset, not the full dataset; while the two releases are structurally similar, results may not transfer directly to the full dataset's broader topical range. Scenario 1 evaluation assumes a single author-verified ground truth per query, which underestimates model capability when several opposing arguments are equally valid counters, so the reported scores should be read as lower bounds. The disputed tag captures a reply author's intent to challenge rather than argumentative substance, so the training and evaluation data contain a mixture of substantive rebuttals and non-substantive conversational exchanges; we do not isolate the contribution of this noise to the observed performance ceiling, and controlled ablations are left to future work. The general retrieval setting (Scenario 2) is formulated with all candidates drawn from the same debate, and our results establish only a lower bound on its difficulty rather than a characterisation of what is achievable with stronger architectures. We evaluate a single bi-encoder backbone (`all-mpnet-base-v2`) and do not explore cross-encoders, larger models, or generation-based approaches. Candidate pool sizes vary across depth levels, and we do not control for this in the depth-based error analysis. Finally, the dataset is English-only and reflects the user population and topical composition of a single platform, which may limit generalisation to other languages and argumentative settings.

Ethical Considerations

The data was collected from a public online debate platform via automated crawling. Because CreateDebate is user-generated and minimally moderated, the full archive necessarily contains debates on extremely sensitive, offensive, or potentially disturbing topics. To balance research utility against the risk of misuse or harm, we adopt a two-tier release: the full raw archive is released only under a gated access agreement restricting use to research

purposes, while the openly released subset is manually curated to exclude broad NSFW categories and (near-)duplicate debates. We deliberately retain all arguments within each selected curated debate including brief, personal, or off-topic replies, to preserve a representative picture of online debate, rather than imposing a post-level filter that could bias the resource. Argument bodies are released as written and may incidentally reference third parties or public figures; we remove user-identifying fields prior to release, and the platform itself is no longer accessible, reducing re-identification risk. The dataset is intended for research on computational argumentation and should not be used to build systems that manipulate, deceive, or harass, or to generate adversarial or inflammatory content. This concern extends to the released retrieval models, which could in principle be used to automate disagreement at scale (e.g., astroturfing or coordinated discourse manipulation); we release them for research on computational argumentation only. Models trained on naturally occurring debate may reproduce social biases and toxic language present in the source data, and any downstream deployment should incorporate appropriate safeguards.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args. me corpus. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 48–59. Springer.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counterargument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1214–1223.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3074–3094.

Dieu-Thu Le, Cam-Tu Nguyen, and Kim-Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical Papers*, pages 1489–1500.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the echo chamber: Detecting countering debate speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7073–7086.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Rudra Ranajee Saha, Laks V. S. Lakshmanan, and Raymond T. Ng. 2024. Stance detection with explanations. *Computational Linguistics*, 50(1):193–235.

Jiasheng Si, Yingjie Zhu, Xingyu Shi, Deyu Zhou, and Yulan He. 2023. Explainable topic-enhanced argument mining from heterogeneous sources. *arXiv preprint arXiv:2307.12131*.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, and 1 others. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, pages 21–25.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.

A Dataset Object Structure

Tables 6 and 7 illustrate the structure and metadata of debate- and argument-level objects.

Table 6: Structure and metadata of a debate object.

Field	Example Value
Title	animal testing
Topic	general
URL	createdebate.com/debate/show/animal_testing
Description	Is it okay to test on animals.
Total Score	195
Total Arguments	83
Total Votes	230
Side: For (Points)	98
Side: Against (Points)	97
Side: For (Tags)	36
Side: Against (Tags)	45

B Additional Dataset Statistics

Figure 2 shows the two structural dimensions of the full dataset discussed in Section 3 but omitted from Figure 1 for readability.

C Data Splits

Tables 8 and 9 report per-split instance counts for the two retrieval scenarios.

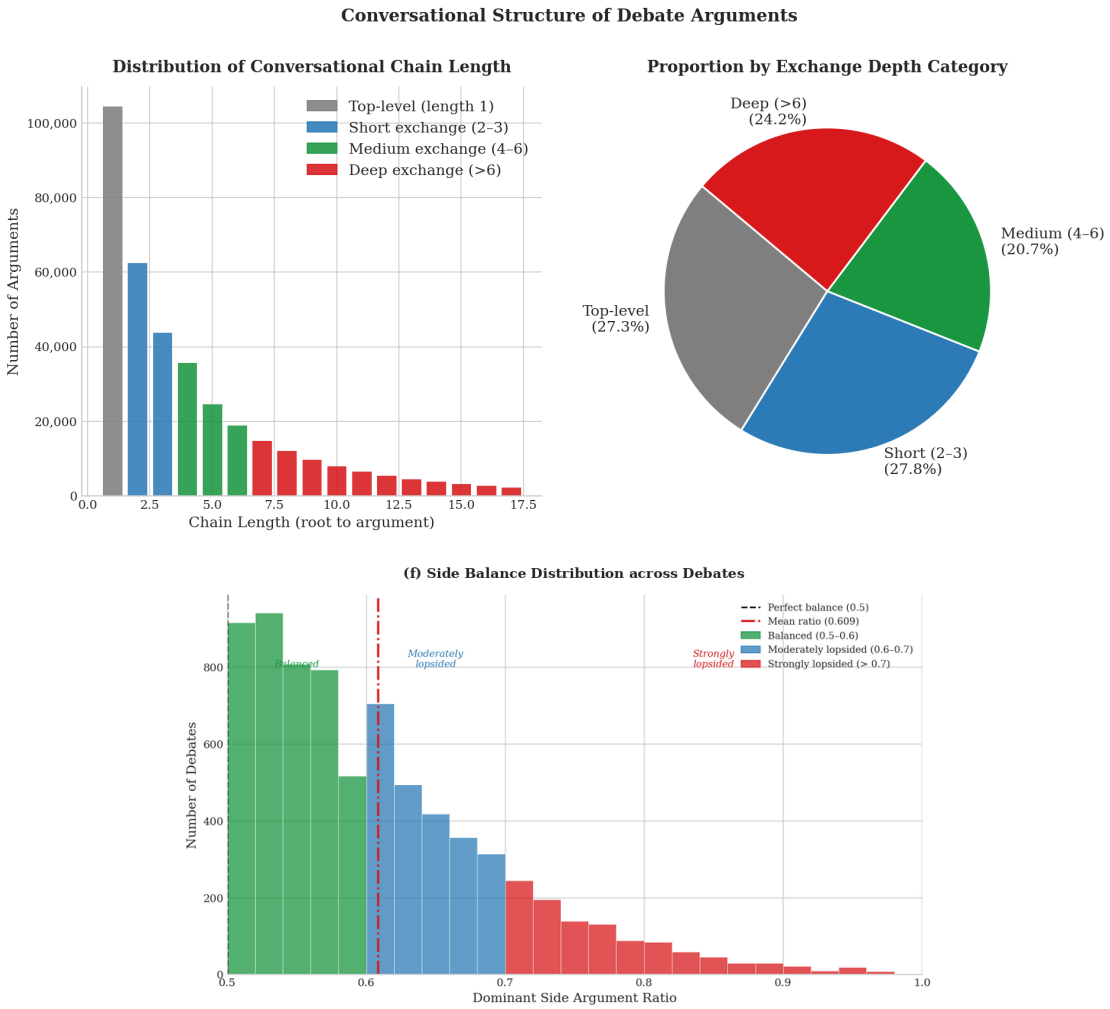


Figure 2: Further structural statistics of the full dataset, continuing the panel labels of Figure 1: (e) conversational chain length and (f) side balance.

Table 7: Structure and metadata of an argument object.

Field	Example Value
Argument ID	arg82671
Parent ID	-1
User Points	6780
Argument Side	For
Argument Points	4
Date	2010-01-15T07:43:03+00:00
Argument Tag	(Empty)
Argument Body	Given the choice between your dog or your mother (or child), who would you choose? I love my cat very much, but I would kill her in an instant to save the life of a loved one. Is animal testing any different?

Table 8: Debate-level data split for Scenario 1 (Targeted Retrieval).

Split	Debates	Queries	Avg. Neg.
Train	456	36,682	119.8
Validation	55	4,009	110.5
Test	56	3,878	93.7
Total	**567**	**44,569**	—

D Qualitative Examples and Error Analysis

This appendix provides the example tables and depth/length figures referenced in Section 4.5. Table 10 shows successful retrievals despite low lexical overlap; Tables 11 and 12 show the two failure types; and Figure 3 shows MRR by query depth and query length.

Table 9: Debate-level data split for Scenario 2 (General Retrieval).

Split	Deb.	Queries	Avg. Pos.	Avg. Neg.
Train	463	65,908	102.6	110.0
Validation	57	7,911	102.9	116.9
Test	57	7,066	83.3	88.7
Total	**577**	**80,885**	—	—

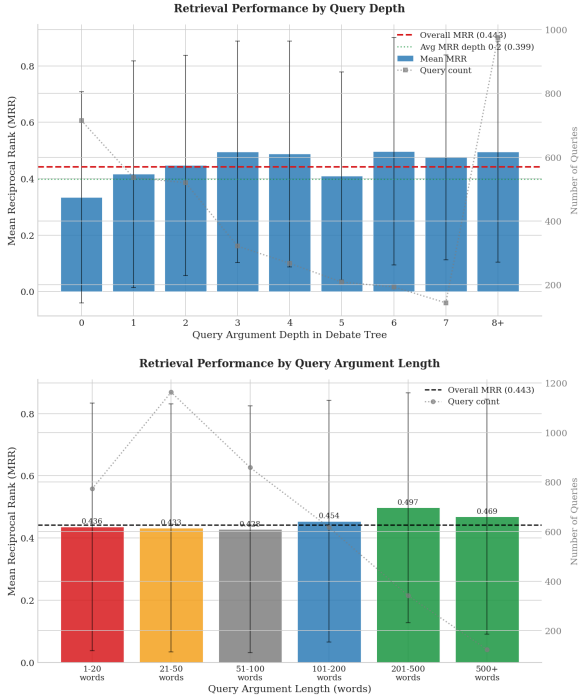


Figure 3: Mean MRR by query depth (top) and by query length in words (bottom) for the fine-tuned bi-encoder on Scenario 1.

Table 10: Success examples: correct counter-argument retrieved at rank 1 despite low lexical overlap.

Query	**Retrieved Counter-Argument (Rank 1 = GT)**
Debate: Is Wal-Mart Good For America? "Personally I would like to say no... but Walmart holds so many jobs right now for Americans." *(Lex: 0.082 | Sem: 0.793)*	"It may hold many jobs for Americans, but have you thought about how many jobs the company has taken away from employing companies overseas? Wal-Mart continuously squeezes its U.S. suppliers to cut their prices [...] the U.S. Bureau of Labor Statistics reported an estimated net loss of 170,000 manufacturing jobs in Ohio in 2005."
Debate: Should the death penalty be used as punishment for specific crimes? "Yes they should, because if they get released they could murder someone again, that's why they should have the death penalty." *(Lex: 0.063 | Sem: 0.776)*	"Just keep them in jail forever so they don't do that again. Don't keep releasing them again and again, just keep them in jail for life."
Debate: TV Violence "TV is one of the biggest resources of information. With TV many people can learn the latest news that happened in the world no matter where they are." *(Lex: 0.152 | Sem: 0.835)*	"There is no argument that TV is useful; the question was 'is it harmful'. TV offers lots of information, but is that info helpful or harmful? Yes, TV gives us the news, but is the news accurate?"

Table 11: Failure examples: pragmatic and contextual rebuttals the model fails to retrieve.

Query	**Ground Truth (GT)**	**Top-1 Retrieved**
Debate: Do gods exist? "[...] They did not feel a higher power, they just had no way to explain their own existence. If you feel a higher power, then I am of the opinion that you should be locked away in a dark, damp place." *(GT Rank: 2 | $Sem_{q,gt}$: 0.619 | $Sem_{q,t1}$: 0.620)*	"Countless people have claimed they've felt a higher power, yet you say they did not feel a higher power [...] If they want to feel something, they feel it."	"In the times before communicating with each other could be done easily, people all over the planet independently believed in a higher power [...] People still wonder about that higher power today."
Debate: Should the government ban the sale of guns? "A guy in China almost did it. He severely injured almost 30 people in a school. Many were close to death." *(GT Rank: 2 | $Sem_{q,gt}$: 0.459 | $Sem_{q,t1}$: 0.459)*	"You just made the best anti-gun argument possible against yourself. Guy in China stabs people, no one dies. Guy with gun in CT, 20 die. Get the difference?"	"No it's the gun. Try killing someone with a finger."

Table 12: Failure examples: non-substantive ground-truth responses.

Query	**Ground Truth (GT)**
Debate: Is god true? A 530-word argument concluding with "You wrote a pargagraph? Wow, you have no life.." *(GT Rank: 97)*	"W.T.H is a pargagraph? Seriously, from a fellow troll... at least have a basic grasp of the English language."
Debate: Is healthcare a right? "This debate is about whether healthcare is a right or not. Stay on topic or leave." *(posted at depth 21)*	"No thanks." *(GT Rank: 135)*

Who Published This Headline? Editorial Fingerprinting in French News Under Topic Control

Amr Sobhy
Le French News Lab
amr@frenchnewslab.org

Abstract

News outlets write in recognizable ways, but how much of that style survives in a single headline once the topic is fixed, and what is it made of? We study outlet attribution from French news headlines under increasingly strict topic control, from category balancing to matched-event comparison. On an ACPM-anchored corpus of 170,160 headlines from ten national outlets across three hard-news categories, with outlets selected through external circulation/audience rankings, a fine-tuned CamemBERT identifies the publisher of an entity-masked headline in a future, unseen year at nearly six times chance; masking outlet self-references barely changes this. We then decompose the fingerprint within the interpretable character model (which alone is surface-strippable). Staged removal of surface form and named entities attributes it, in order-dependent shares, largely to presentation (casing, punctuation, typographic house style) and to a lexical residual that survives removing both, with a smaller share carried by which entities are named. An interpretable tone family adds almost nothing at the single-headline level, so the residual is not reducible to tone here. Category control alone cannot separate editorial form from an outlet's story selection, so as a final check we hold the reported event constant, clustering same-event headlines across outlets and comparing them within cluster with a human-validated fixed-effects estimator. Both pre-specified attribution contrasts fall to near zero, while one validated difference survives: BFMTV foregrounds institutional and corporate actors as grammatical agents more than its peers. The result shows how much control is required before outlet differences can be read as framing rather than agenda. We release identifiers, hashes, code, derived features, the framing codebook, and gold annotations.

1 Introduction

News outlets write in recognizable ways. Beyond the events they choose to cover, publishers differ in tone, register, punctuation, capitalization, and the small typographic conventions of a house style. We ask how much of this editorial identity is recoverable automatically, and from how little text: can the publishing outlet be predicted from a single headline, once the topic is held constant? The question has practical stakes for web-derived corpora, which rarely preserve reliable source metadata: a recoverable outlet style bears on corpus provenance and source attribution, near-duplicate and syndication detection, and auditing register bias in scraped news collections.

We call this task *editorial fingerprinting*, and it connects two literatures that have largely stayed apart. Stylometry and authorship attribution recover the author of a text from measurable stylistic regularities (Stamatatos, 2009), but typically operate on long documents and individual authors rather than institutions. Work on news sources, in turn, predicts an outlet's factuality or political bias from article bodies and external metadata (Baly et al., 2018), or studies genre and subjectivity in the press (Escouflaire et al., 2024); it rarely asks whether the outlet itself is identifiable from the surface form of a single item. For French in particular, computational study of the media ecosystem has focused on outlet positioning and sentiment rather than attribution (Dufraisse et al., 2024).

Outlet attribution is easy to get wrong, because outlet identity is heavily confounded with topic: an unchecked classifier will recover the outlet from its subject matter rather than its style. Our design reduces agenda confounding through category balancing, entity masking, and, for framing, matched-event comparison. We assemble a category-controlled corpus of headlines from ten major French national outlets, balanced across

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 22–36
October 29, 2026 ©2026 Association for Computational Linguistics

three hard-news categories, with outlet selection anchored on external circulation/audience rankings so the panel is auditable. We evaluate a ladder of models of increasing capacity and decreasing interpretability to locate the attributing signal rather than merely maximize accuracy, and test every model under outlet self-reference masking, named-entity masking, and a temporal split that trains on past years and tests on a future, unseen year.

Our contributions are threefold. First, we introduce a topic-controlled benchmark and a reproducible, externally anchored outlet-selection methodology for French headline attribution. Second, we *decompose* the editorial fingerprint, within the interpretable character model, into presentation, named-entity, and lexical-residual components through staged surface-stripping and entity masking, and show the residual is *not* tonal at the single-headline level. Third, we stress-test category-level control with a matched-event design: holding the reported event constant, most apparent attribution differences attenuate, while one narrow, gold-validated component of editorial identity, grammatical agency, survives. The throughline, separating *how* an outlet writes from *what* it covers, is rigor and a reusable resource rather than a new model.

2 Related Work

Our task sits at the intersection of three lines of work: authorship attribution and stylometry, source- and outlet-level prediction over news text, and French-language news NLP.

Attributing a text to its author from measurable stylistic regularities is a foundational NLP problem (Stamatatos, 2009). The dominant features (character n-grams, function-word and punctuation distributions, lexical features, usually with linear SVMs) capture the unconscious habits of an author or the mandated conventions of an editorial desk. Character-level features are especially valued for punctuation, casing, and orthographic regularity and for noise tolerance (a natural fit for headlines), though as Stamatatos notes they also encode lexical and contextual cues, not style alone. The tradition degrades sharply on short text, but has been pushed to provenance-oriented settings such as machine-generated-news detection, where the signal tracks *provenance* rather than veracity (Schuster et al., 2020). We adapt it to the extreme short-text regime of the headline, contrasting word- and character-level representations to isolate which carries outlet identity.

A parallel line treats the *news source* rather than an individual author as the target. Baly et al. predict factuality and political bias of entire media organizations from article samples and metadata (Baly et al., 2018), establishing that outlet-level signal is real but leaning on features beyond the text (Wikipedia, Twitter, URL and web-traffic cues) and operating at the level of *organizational reputation*, not the surface form of one item. We instead predict the outlet from the *text of one headline alone*, category-controlled and entity-masked so the analysis reduces, but does not eliminate, agenda-related signal, and (where this literature treats source identity as an artifact to suppress) quantify it directly. Our corpus construction also draws on the web-as-corpus tradition, where boilerplate and template stripping are long-standing prerequisites for web-derived corpora (Baroni et al., 2009; Pomikálek, 2011; Schäfer and Bildhauer, 2013); but where that work treats residual markup as noise to minimize, we treat it as a measurable leakage channel and audit its contribution to the attribution signal directly.

Between sparse linear stylometry and large transformers sit character-level convolutional networks, which classify text directly from raw characters without word-level knowledge (Zhang et al., 2015). Operating on the character stream makes CharCNNs a natural candidate for signal carried by punctuation, casing, and orthographic house style. We include one as a mid-point on a representational ladder, so accuracy gains can be attributed to level of representation rather than capacity alone.

CamemBERT and the contemporaneous FlauBERT brought high-quality monolingual pretraining to French (Martin et al., 2020; Le et al., 2020). Applications to the French press concentrate on *subjectivity and genre* and on *disinformation*. For genre, Escouflaire et al. find a fine-tuned CamemBERT beats a 32-feature model on opinion-versus-news classification while noting the feature model stays competitive and the transformer costs computation and explainability (Escouflaire et al., 2024); for disinformation, OBSINFOX annotates unreliable French sources (Icard et al., 2024). Closest to ours, Dufraisse et al. map the French media ecosystem by combining objective and subjective analyses of political

news with CamemBERT (Dufraisse et al., 2024), but target sentiment and outlet positioning, not attribution. To our knowledge, outlet attribution from the headline alone, under explicit topic control and temporal and entity-masking ablations, has not, to our knowledge, been studied in French; we address this gap.

Finally, our strictest analysis connects attribution to the communication-studies tradition of *framing*: making some aspects of a perceived reality more salient (Entman, 1993), and how the press frames political issues through presentation rather than content (Iyengar, 1991). Such framing is carried by grammar as much as word choice: syntactic packaging like voice and agentive-subject assignment conveys implicit stance even with lexical content fixed (Greene and Resnik, 2009). We operationalize this with parser-based measures of agency, voice, and attribution and, crucially, hold the reported *event* constant via a matched-event design, which substantially reduces event-level agenda confounding, letting residual differences be read as framing under the stated instruments.

3 Data

We study outlet attribution over ten major French national outlets, restricted to three hard-news categories—International (international affairs), Politique (political news), and Économie (economic news)—to reduce agenda confounding. To keep selection auditable rather than ad hoc, we anchor the selection criterion on ACPM (Alliance pour les Chiffres de la Presse et des Médias), France's official press-measurement body: a generalist national daily with continuous 2021–2025 hard-news coverage and sufficient per-category volume. The core is the six ACPM-ranked national dailies (*Le Monde, Le Figaro, Les Échos, Libération, La Croix, Le Parisien*); for format diversity we add three national broadcast newsrooms (*Franceinfo, BFMTV, TF1 INFO*), and for regional diversity *Ouest-France*, France's largest-circulation daily. From the ACPM daily rankings we exclude only ineligible titles: *L'Équipe* (sports-only), the youth dailies *Le Petit Quotidien* and *Mon Quotidien*, the *New York Times* international edition (foreign), and *L'Humanité*, excluded for a corpus-collection shortfall against our per-category floor, not an editorial judgment. The study uses the articles available in an existing collection of the outlets' online editions rather than claiming an exhaustive crawl of everything they published. Each record supplies its publisher, publication date, canonical site category, URL, and headline extracted from page metadata, with an outlet-specific CSS selector used as a fallback when necessary. We restrict the collection to 2021–2025 and the three categories below, normalize the stored headlines, remove malformed and non-article records and exact duplicates, and then draw the balanced sample used for modeling.

Headline cleaning. To reduce web-collection artifacts before measuring editorial differences, we clean each as-published headline before modeling, in all conditions and before masking: stripping templated site-name suffixes (e.g. a trailing "$-$ *Outlet*"), leading broadcast/format tags (VIDÉO, EN DIRECT, REPLAY), residual URLs and separators, and non-article index pages. The cleaning is conservative and verified: it alters 2.5% of headlines overall (concentrated in broadcasters; $\leq$ 2% for print), and every sampled edit removes only presentation, never editorial content.

We balance every outlet$\times$category cell at 5,672 headlines (170,160 total; 17,016 per outlet, 56,720 per category; Table 1), using headline text alone. We report pooled accuracy over the balanced mixture of all three categories and separate models trained and tested within each category. The latter test whether the publisher remains predictable without relying on differences among International, Politique, and Économie coverage. We evaluate under two regimes: a random split (17,016 test headlines) and a temporal split training on 2021–2024 and testing on the unseen year 2025 (35,080 headlines), which is our primary setting. For copyright reasons, we release article identifiers, content hashes, extraction and sampling code, derived features, and per-headline category labels.[1] These materials support attempted reconstruction of the balanced benchmark from public sources and record-level checking against the original hashes. The released code regenerates the figures and prediction-based attribution tables from text-free derived features, while rederiving the character-level tables requires reconstructing the balanced corpus. The matched-event framing analysis (Section 5.4) draws instead from the full article pool (478,110 titles), restricted to

[1] All released artifacts (code, derived features, framing codebook, gold annotations, and the corpus-rebuild manifest): https://github.com/lefrenchnewslab/who-published-this-headline

Stratum	Outlets	Source
National dailies	Le Monde, Le Figaro, Les Échos, Libération, La Croix, Le Parisien	ACPM PQN
Broadcast	Franceinfo, BFMTV, TF1 INFO	Format-diversity addition
Regional	Ouest-France	ACPM PQR

Table 1: Corpus composition (N=10). Three categories $\times$ 5,672 per outlet$\times$category = 170,160. Train 2021–2024, test 2025 (35,080).

the same categories and panel and cleaned identically, recovering events as cross-outlet clusters of near-duplicate, same-window headlines; its released result tables are verifiable, but the unshipped full-pool input prevents independent reconstruction from the supplement alone.

4 Method

The study has four stages. First, we construct a category-balanced headline corpus and define random and future-year evaluation splits. Second, we compare model families under raw, outlet-masked, and entity-masked inputs to measure publisher predictability under progressively stronger controls. Third, we strip presentation and entity information from the interpretable character model to examine which signals sustain its predictions. Fourth, we compare outlets within clusters of headlines about the same event, where differences can be interpreted more narrowly as framing under the validated measures. The word model tests lexical choice, the character models add subword and presentational conventions, and the contextual encoders test what richer representations recover beyond those interpretable baselines; XLM-R separately tests whether the result depends on French-specific pretraining.

We frame outlet attribution as single-label classification: given one headline, predict its publishing outlet. The label is the publisher recorded at collection, so unlike annotated-topic benchmarks the target carries no annotation noise or subjective label validity to audit. To locate where the attributing signal lives, we evaluate a ladder of models of increasing capacity and decreasing interpretability, from sparse lexical features to a fine-tuned transformer, against a most-frequent-class baseline. The two linear models isolate word- versus character-level signal. The word model is an L2-regularized logistic regression over TF–IDF word unigrams and bigrams (up to 120K features, sublinear term frequency, minimum document frequency 2). The character model is a linear SVM over TF–IDF character n-grams of length 3–5 with word boundaries (up to 200K features). Both use balanced class weights and fixed regularization (logistic regression $C = 4$, SVM $C = 1$). Between the linear models and the transformer we train a character-level convolutional network in the style of Zhang et al. (2015): a six-layer 1-D convolutional tower (64–256 channels, ReLU, max-pooling) over a learned character embedding, then a two-layer classifier with dropout; a hybrid variant adds mean-pooled word embeddings. Headlines are encoded as their first 200 characters (50 words for the hybrid); training uses Adam (5×10^{-4}, weight decay 10^{-4}, batch 64, cross-entropy) for up to 30 epochs with early stopping (patience 3) on macro-F1. At the top of the ladder we fine-tune CamemBERT-large (Martin et al., 2020) (335M parameters) with a classification head (max length 64, batch 256, weight decay 0.01, 10% warmup, label smoothing 0.1, up to 10 epochs, patience 2 on macro-F1). The learning rate (1×10^{-5}) was selected by a prior sweep over $\{5, 2, 1\} \times 10^{-5}$ and fixed across all conditions; per-condition tuning could shift absolute numbers but not the ablation pattern. The 2025 test year informed no hyperparameter, threshold, or design choice; all model selection used only 2021–2024 data. To test whether attribution leans on French-specific pretraining, we also fine-tune the multilingual XLM-R-large (Conneau et al., 2020) ($\sim$550M parameters) in the primary configuration only (temporal split, entity-masked, three seeds: 42, 13, 1234), holding data, masking, and hyperparameters identical so the sole change is monolingual French versus multilingual pretraining.

To separate editorial style from confounds, we run every model under three text conditions. In the **raw** condition the headline is used as published. In **mask-outlet**, outlet self-references (name variants and domains) and URLs are replaced with placeholder tokens, removing the trivial cue of an outlet naming itself. In **mask-entities**, we additionally replace person, location, organization, and miscellaneous entity spans, detected with spaCy's French NER model, with a placeholder; this tests

whether attribution survives the removal of *who and what* a headline is about, leaving primarily stylistic form. Under the two evaluation regimes of Section 3, linear models and the baseline are deterministic and run with a single seed; the Char-CNN, CamemBERT, and XLM-R are run with three seeds (42, 13, 1234) on the temporal headline configurations, and we report the mean and standard deviation. We measure accuracy, macro-F1, and top-2 accuracy, and report bootstrap confidence intervals and McNemar tests on the temporal setting to establish that differences between models are significant.

4.1 Matched-event framing

The attribution analyses above hold topic constant only at the category level. To read *framing*, we hold the reported event constant and compare how different outlets phrase the same story.

Recovering events. Working on the full article pool (Section 3), we embed each headline with a French sentence encoder (sentence-CamemBERT-large) and, within each (category $\times$ calendar-month) block, run community detection at cosine similarity ≥ 0.70. A cluster is admitted as a *matched event* only if its headlines fall within a $\pm$2-day window, share at least one named entity, and come from at least four distinct outlets; the 0.70 cosine cutoff is not a claim that this similarity equals same-event identity but a proxy set a priori whose same-event precision is 0.94 on a 400-pair hand-labeled sample, with cluster purity additionally spot-checked on a 30-cluster audit. This yields 626 matched events spanning 3,006 headlines and all ten outlets (per-outlet coverage is uneven, from Libération's 402 to Ouest-France's 111, with BFMTV at 290).

Framing instruments, validated against humans. From a Universal-Dependencies parse and curated lexicons we extract six headline-level framing measures: *valence* (a validated sentiment ensemble), *attribution* (reporting and stance verbs governing a clausal complement or quote, inverted quotatives, and *selon*-type ("according to") markers), *passive* voice, *agent_subject* (an agentive subject performing the main action), *verbless* construction, and *hedging*. The exact rules are frozen in a released codebook. Crucially, a measure is admitted to the confirmatory analysis only if it clears a pre-set agreement gate against 240 headlines hand-coded by an expert annotator from the codebook alone, blind to the extractor output: Cohen's $\kappa \geq 0.60$ for the binary measures and Spearman $\rho \geq 0.50$ for valence. Five of the six pass (attribution $\kappa = 0.97$, hedging 0.84, passive 0.75, agent_subject 0.63, valence $\rho = 0.62$); verbless, a deterministic parse property, was not gold-validated and is reported as exploratory only. The confirmatory framing results thus rest on validated instruments, though residual measurement error may attenuate small effects. A second expert independently re-coded the 240 headlines (blind to the first rater and extractor); human–human agreement is high (attribution/hedging $\kappa = 0.92$, agent_subject/passive 0.85, valence $\rho = 0.79$), so the constructs are reliably codable; agent_subject's instrument–human agreement is lower ($\kappa = 0.63$).

Within-event estimator. For each headline, we subtract the average value among headlines about the same event. The resulting comparison asks whether an outlet systematically departs from the wording shared for that event rather than whether it covers different events. We adjust for headline length, and the uncertainty estimates account for headlines grouped within the same event. We pre-specify four directional contrasts derived from the attribution and tone analyses (BFMTV and Les Échos on attribution, BFMTV on agency, Libération on valence) and test them one-sided with Bonferroni correction (measures, estimator, and directional priors fixed before the confirmatory fits; commit-timestamped provenance in Supplementary S1); all other outlet$\times$measure cells are reported as exploratory under Benjamini–Hochberg FDR. With 626 events the design detects within-event effects of about 0.12 at 80% power; binary measures estimated over all 3,006 headlines are recovered with greater precision, as for agency.

5 Results

We present the results as a single descent under increasingly strict topic control (Acts I–IV below). Acts I–III (benchmark, decomposition) stand on their own; the matched-event probe (Act IV) is a strict confirmatory add-on.

5.1 A fingerprint exists, and it is uneven

Table 2 reports outlet-attribution accuracy for every model, split, and text condition (chance $1/10 = 0.100$). In the primary setting (a temporal split training on 2021–2024 and testing on the

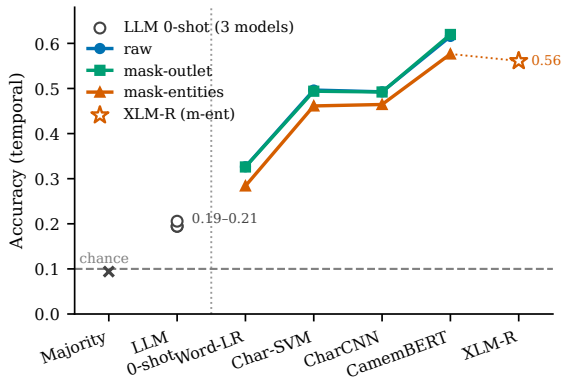


Figure 1: Accuracy ladder (temporal split). The trained ladder rises from Word-LR to CamemBERT; the largest gain is the word→character step, outlet masking barely moves accuracy, and entity masking costs a few points at each rung. Three zero-shot LLMs (`gpt-oss-120b`, `Llama-3.3-70B`, `Mistral-Small`, left of the divider) cluster just above chance and below every trained model; the multilingual XLM-R (star) lands just below CamemBERT. Chance $= 1/10$.

unseen year 2025), a fine-tuned CamemBERT recovers the publisher of an *entity-masked* headline at 0.576 accuracy (macro-F1 0.569; mean over three seeds, 95% CI $[0.573, 0.584]$), roughly $5.8\times$ chance and above every weaker model; on raw text it reaches 0.615 temporally and 0.643 on the random split (top-2 0.80). Each of the three released primary-cell comparisons is significant (McNemar $p < 10^{-300}$).

The accuracy ladder localizes the signal (Figure 1). The word-to-character step is the largest single jump (temporal raw: 0.326 to 0.496, +17 points), showing that character-level information—including orthography, casing, punctuation, and lexical subword patterns—is more predictive than word features alone. The character CNN does not surpass the linear character SVM temporally (0.493 vs 0.496) and only matches it on the random split, so the character signal is captured by n-grams rather than architecture. CamemBERT adds a further 12 points, quantifying what contextual modeling contributes over surface character statistics. A zero-shot probe tests whether outlet identity comes from parametric world knowledge rather than learned corpus regularities. Prompted with the ten outlets on a raw 2025 headline, three open-weight LLM families[2] recover the publisher at only 0.19–0.21 accuracy (1.9–$2.1\times$ chance), far under even the word model (0.326) and collapsing onto a few outlets (macro-F1 0.17–0.19); five demonstrations per outlet do not help (Appendix D). The recoverable signal is thus primarily corpus-specific surface regularity rather than knowledge elicited by prompting alone: a no-training floor, not a tuned baseline. Nor does the fingerprint depend on French-specific pretraining: multilingual XLM-R-large, fine-tuned identically, recovers the entity-masked publisher at 0.561 ± 0.001 across three seeds (Table 2), within 1.5 points of CamemBERT.

The signal is robust to the confounds the design targets. Masking outlet self-references does not reduce accuracy (CamemBERT temporal 0.615 raw vs 0.620 masked); the small difference if anything favors the masked condition (McNemar $p = 0.35, 0.001, <0.001$ over seeds 42/13/1234), so the model does not rely on outlets naming themselves. Masking person, location, and organization entities costs about four points (0.615 to 0.576), so attribution does not depend heavily on which actors a headline reports. This masking is not a near-identity: it fires on 87.5% of headlines and recovers 97% of gold spans (150-headline sample). Within a single category the publisher stays well above chance (Table 3): 0.61 Politique, 0.58 Économie, 0.54 International, each 5.4–$6.1\times$ chance. Two checks reduce broad category-level confounding but do not eliminate event or beat information: with every cell balanced, a category-only classifier is at chance, and a model trained *and* tested inside one category still recovers the outlet at $\sim 5\times$ chance (char-SVM 0.48). Removing boilerplate leaves accuracy unchanged to four decimals, and the top recurring lead is under 0.4% of any outlet, providing no evidence that recurring templates explain the result in this audit.

Fingerprintability is highly uneven across outlets (Figure 2; Appendix Figure 6). Distinctive registers are recovered strongly: Ouest-France (F1 0.76, precision 0.87), BFMTV (0.75), Franceinfo (0.62), whereas the generalist Paris dailies are mutually confusable, Le Figaro hardest (0.35), collapsing toward Le Parisien. Some newsrooms thus hold a sharply identifiable house style, others an interchangeable national-daily register. The most- and least-fingerprintable outlets are stable across seeds and all ten fixed-model leave-one-outlet rescoring (per-outlet F1 rankings at Spearman $\rho \geq$

[2] `gpt-oss-120b`, `Llama-3.3-70B`, and `Mistral-Small`.

	Random			Temporal		
Model	raw	m-out	m-ent	raw	m-out	m-ent
Majority	.100	.100	.100	.094	.094	.094
Word-LR	.361	.358	.304	.326	.326	.284
Char-SVM	.527	.523	.479	.496	.494	.461
CharCNN	.531	.526	.489	.493	.492	.464
CamemBERT	.643	.646	.596	.615	.620	**.576**
XLM-R†	—	—	—	—	—	.561

Table 2: Outlet-attribution accuracy by model, split, and text condition. Chance = 0.100. Bold = headline model (CamemBERT, temporal, entity-masked). Neural temporal cells: mean over three seeds (std $\leq$ 0.02; 0.003 for the bold cell); others single-seed or deterministic. †XLM-R-large: multilingual comparator, primary cell only, mean over three seeds (std 0.001), testing whether attribution depends on French-specific pretraining. Three decimals are kept because some masking contrasts (e.g. 0.615 vs. 0.620) differ by under 0.01; that outlet-masking delta does not favor the raw condition (McNemar, §5), so it is not evidence of reliance on self-naming.

Category	Accuracy	n
Politique	.61	11463
Économie	.58	11728
International	.54	11889

Table 3: Within-category accuracy (CamemBERT, temporal, entity-masked; seed 42). n = per-category test size under temporal split. Chance = 0.100. Category ordering is stable across seeds (accuracy varies $\leq$ 0.008).

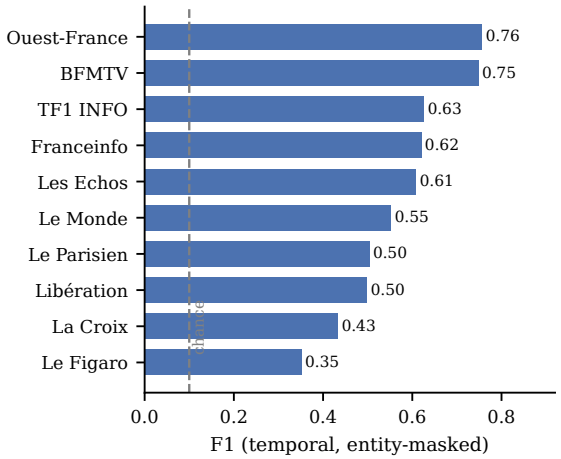


Figure 2: Per-outlet F1 (CamemBERT, temporal, entity-masked; seed 42), sorted. Uneven: Ouest-France (0.76) to Le Figaro (0.35); regional, business, and broadcast outlets lead; generalist Paris dailies trail. Chance = 1/10.

0.93; Appendix B).

5.2 What the fingerprint is made of

Of what is this signal composed? We bound it from above (naming the surface conventions that carry it) and from below, stripping surface form entirely to measure what survives. Inspecting the most discriminative character features points to a surface-convention account of house style, spanning typographic habits and editorial-desk labels such as live/video markers and agency credits, conventions of presentation rather than orthography, listed per outlet in Table 4.

To test whether the fingerprint reduces to these surface markers, we normalize them directly (mapping guillemets and curly quotes to straight quotes, curly apostrophes to straight, and rewriting ellipses, dashes, and trailing terminal punctuation to a canonical form) and re-fit the character SVM on the temporal split. Accuracy falls only from 0.496 to 0.427 (a 14% relative drop): the named typographic conventions carry real signal but a minority of it, and the large majority of the character-level fingerprint survives normalization, reflecting deeper regularities of casing, spacing, and sub-word form rather than a handful of formatting tokens. We next ask how much of the fingerprint a compact, nameable set of dimensions can explain. We compute twenty interpretable headline features: typographic (guillemets, ellipses, colons and colon-led topic prefixes, question and exclamation marks, percent signs, dashes), structural (character and word length, mean word length, digit density, listicle openings, all-caps tokens, capitalization ratio), and register (function-word ratio), and fit a multinomial logistic regression on these features alone, on the same temporal split. With no lexical or sub-word information it reaches 0.269 accuracy (2.7× chance; top-2 0.44), recovering 43% of the char-SVM's above-chance margin. The per-outlet profile (Figure 3; released with the code) makes the house styles legible: Ouest-France headlines are short, capital-initial

Outlet	Salient discriminative markers
Le Monde	« », curly apostrophe, “;”, %
Le Figaro	« » with trailing colon (»:)
Les Échos	« », colon, apostrophe
Libération	« » (multiple variants)
La Croix	colon-led “*X*:”, agency credit
Le Parisien	ellipsis “…”, « », comma
Franceinfo	straight quotes ", colon
BFMTV	colon “*X*:”, question “?”
TF1 INFO	pipe “|”, colon, “direct”
Ouest-France	period-terminated headlines (e., n.)

Table 4: Most discriminative character features per outlet (char-SVM, raw text): typographic conventions and editorial-desk labels; markers are post-cleaning (non-leading positions). Top features per outlet released.

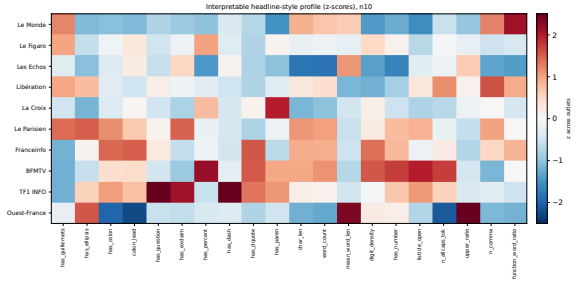


Figure 3: Per-outlet z-score profiles across 20 interpretable headline features (typographic, structural, register). Each cell shows the outlet's standardised deviation from the corpus mean. Guillemets separate print dailies from broadcasters; period-termination (Ouest-France) and colon-led constructions (BFMTV, TF1 INFO, Franceinfo) are the strongest positive signatures.

statements rarely carrying a colon-led topic prefix (12%, versus 43–56% for the broadcasters), while guillemets (the single most heavily weighted feature) cleanly separate the print dailies (9–33%) from the broadcasters (0%). A small inventory of presentational conventions thus accounts for a substantial share of the surface signal, the remaining gap reflecting finer sub-word regularities that resist naming.

We now bound it from *below*: how much survives once surface form is removed. Re-fitting the character SVM on the temporal split and stripping each headline in stages, lowercasing and deleting all punctuation, digits, and whitespace collapses accuracy from 0.496 (raw) to 0.324 (3.2× chance), and additionally masking every spaCy-NER span lowers it to 0.270 (2.7× chance). A word-level logistic regression on the same stripped text tracks the character model to within 0.002 at every stage, so what remains is genuinely lexical rather than a sub-word artifact. This yields a clean three-way decomposition of the *char-SVM's* raw above-chance margin: roughly 44% presentation (typography, casing, numerals), 14% which entities are named, and 43% a lexical residual surviving removal of all surface form *and* all named entities (95% bootstrap CIs in Figure 4). These shares are ablation-defined and order-dependent, not an additive causal partition, and specific to the interpretable character model; the transformer is not surface-strippable in stages, but its entity component matches, losing a comparable margin share to entity masking ($0.615 \rightarrow 0.576$ vs. the char-SVM's $0.496 \rightarrow 0.461$; Table 2). Word choice alone thus recovers the outlet at nearly three times chance. The residual's most distinctive words show its cleanest style-coded slice is attribution language (*BFMTV*: *estime/assure/affirme que*, “states/asserts that”), but it stays entangled with beat vocabulary (*évêques* “bishops” for La Croix, *cryptos* for Les Échos, *commune* for Ouest-France), tracking what each outlet covers as much as how it writes. Stripping the top outlet-distinctive content words nonetheless closes under 40% of the residual's above-chance margin, leaving a 2.3×-chance residual after this tested blocklist (Appendix F). Isolating the framing component requires holding the story constant (Section 5.4).

5.3 The residual is not (just) tone

An obvious candidate for the residual is tone: differences in sentiment or stance. We compute an interpretable *tone* family: sentiment polarity and intensity (a French sentiment model and a polarity lexicon), subjectivity, and hedging, attribution, and sensational-marker rates. On the same temporal split these alone recover the outlet only weakly (0.13, 1.3× chance), and adding them to the twenty surface features lifts accuracy by just 0.004 pooled and 0.010 within category, every tone feature ranking below the typographic block by weight. The gain is larger within category than pooled, so the small tonal signal is editorial, not topical. A single headline thus encodes editorial identity through presentational form far more than measurable tone. Because our headline-length sentiment proxies are noisy ($\rho = 0.33$ between backends), this bounds *this* tone family, not richer stance models.

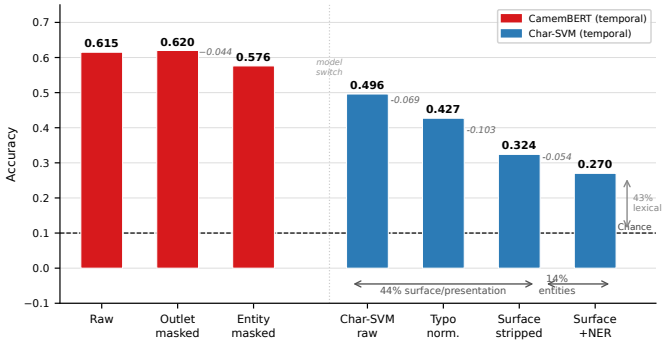


Figure 4: Signal-stripping waterfall (temporal split). *Left, red:* CamemBERT under progressive masking; *right, blue:* Char-SVM as surface form is stripped in stages (the two halves use different model families, not a single continuous descent). Brackets split the char-SVM above-chance margin (0.396): 44% presentation, 14% named entities, 43% lexical residual; the split is ablation-defined and order-dependent, not an additive causal partition.

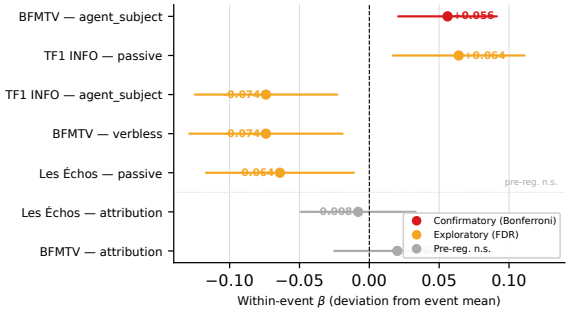


Figure 5: Within-event framing contrasts (β = deviation from event mean; within-event estimator, Section 4.1; 626 events, 3,006 headlines). Red: survives Bonferroni ($p_{\text{Bonf}} = 0.008$); orange: exploratory (FDR < 0.05); grey: pre-specified, n.s. (Libération valence $\beta = -0.021$, $p_{\text{Bonf}} = 1.0$).

5.4 What survives matched-event control?

This final act is a deliberately conservative test: with pre-specified contrasts, gold-validated instruments, and Bonferroni correction, attenuation is the expected outcome and the finding is which single difference clears the bar. Under matched-event control, the only pre-specified contrast that survives is grammatical *agency*; attribution contrasts attenuate toward zero (Figure 5). On the same event, BFMTV assigns an agentive subject markedly more often than the average outlet ($\beta = +0.056$; one-sided $p = 0.002$, $p_{\text{Bonf}} = 0.008$); these agentive subjects are predominantly institutional and corporate actors (see Limitations).

Exploratory cells trace a coherent fingerprint on the same axis (all FDR-significant): TF1 INFO is less agentive ($\beta = -0.074$) and more passive ($+0.064$), Les Échos less passive (-0.064), BFMTV less verbless (-0.074; not gold-validated). Broadcasters thus write actor-driven headlines while TF1 INFO leans to passive, event-centered ones for the same story.

Within the same event, both pre-specified attribution contrasts fall to near zero (BFMTV $\beta = +0.020$, $p_{\text{Bonf}} = 0.20$; Les Échos $\beta = -0.008$, $p_{\text{Bonf}} = 0.76$). So agency is an event-controlled *framing* difference while attribution tracks outlet *beat* composition. The result is robust to syndication removal, leave-one-outlet-out, and a 626-event jackknife (Appendix E).

6 Conclusion

French headlines carry a substantial publisher signal even after category balancing, entity masking, and evaluation on a future year, although these controls do not eliminate event and beat information. Within the interpretable character model, presentation and lexical choices both contribute to this source predictability. The stricter same-event analysis supports a narrower conclusion: three of four pre-specified framing contrasts approach zero, while BFMTV more often foregrounds institutional and corporate actors as grammatical agents. Editorial fingerprinting can therefore diagnose source and register effects in web corpora, while claims about framing require direct control for the reported event. Future work should test whether these patterns transfer across languages, outlet types, and longer periods.

Limitations

Our study is limited to ten generalist French national hard-news outlets and to headlines; we do not test whether the findings extend to other languages, to local or specialized press, to soft-news

categories, or to full article text. Topic is held constant by category, and category assignment may be imperfect, so a small amount of residual topic signal may remain; the within-category results (Table 3), where the outlet is recovered at 5.4–$6.1\times$ chance inside each single category, directly bound how much the pooled numbers could owe to cross-category topic leakage. Some of the recovered signal reflects recurring editorial-desk and section conventions rather than purely linguistic style, and our outlet-name masking is pattern-based and may not catch every self-reference. Entity masking relies on automatic NER, whose span recall on a hand-annotated headline sample is high ($\approx$97%) but not perfect, so a small amount of residual entity signal may persist in the entity-masked condition. We operationalize house style as classifier-recoverable surface form and do not validate the broader notion of house style itself against human judgments, although the framing instruments used in Section 5.4 are human-validated. The signal is carried more by presentation than by word choice, but a lexical residual persists once all surface form and named entities are removed; because that residual remains entangled with each outlet's beat vocabulary, we isolate its framing component only through the matched-event design of Section 5.4. That design is a focused confirmatory probe rather than the backbone of the paper: the benchmark and decomposition of Acts I–III rest on the full 170,160-headline corpus and stand independent of it. The matched-event analysis is necessarily smaller (626 events over 3,006 headlines) but we guard it accordingly: instruments are admitted only after passing a gold-validation gate, the four contrasts are pre-specified, standard errors are clustered by event, clusters require at least four distinct outlets, and the result is stable under a cluster-purity check, a syndication (wire-copy) refit, a cluster-membership sweep, and a headline-length-cap sweep (Appendix E). The design is powered for moderate within-event effects (n needed 133–143 versus 626 available), and the agency effect, estimated as a binary measure over all 3,006 headlines, is recovered with precision: its binary within-event estimator has a minimum detectable effect of 0.060 at 80% power, half the 0.12 continuous-measure figure. The observed 0.056 is slightly below this nominal MDE (attained power ≈ 0.73 at the Bonferroni threshold; ≈ 0.89 uncorrected) (Appendix E). Scaling the clustering to the full article pool to grow the event count is a natural next step. Of the four pre-specified confirmatory contrasts only one (BFMTV agency) survives Bonferroni correction, so we advance agency as a specific, validated, topic-invariant framing signal rather than a claim that outlets differ wholesale in framing; the remaining exploratory cells, while coherent, await confirmation. Matched-event coverage is uneven across outlets, so power is not uniform, and clustering recovers only the subset of stories covered by at least four outlets. Finally, the agency instrument has the lowest instrument–human agreement of our measures ($\kappa = 0.63$): against the pooled 240-headline gold set its precision is 0.89 at recall ≈ 0.66, so it *under*-detects agentive subjects in aggregate. This would attenuate the within-event BFMTV estimate under non-differential under-detection; differential extraction error remains a limitation. The excess is concentrated in institutional and corporate agents (ministries, companies, states); restricting the subject to an animate human referent leaves too few within-event instances to resolve the contrast ($\beta \approx +0.02$, n.s.). We read the effect as *institutional-actor framing*, not a claim about individuals, and scope the abstract and Section 5.4 accordingly. Temporal robustness is assessed on a single held-out year (2025), so longer-horizon drift is not measured. Finally, the copyright-driven release of derived artifacts rather than raw text constrains replication that needs the raw headlines themselves, though the anonymized code and derived features, released as supplementary material, contain every reported table and figure and regenerate all figures and the prediction-based attribution tables without raw text, while re-deriving the character-level tables requires rebuilding the balanced corpus from the released identifiers; the framing tables are verifiable from released results but not independently reconstructible without the unshipped full-pool input. We did not measure what fraction of the manifest remains recoverable from live pages or public archives, and therefore make no complete-reconstruction claim. A reconstruction should report unavailable pages and hash mismatches as unrecovered records rather than silently substituting changed content.

Ethics Statement

This work attributes published news headlines to their publishing organization; it profiles institutional editorial style, not individuals, and uses

only publicly published material. We see limited risk of harm: the task could in principle aid deanonymization of an outlet that wished to publish anonymously, but the outlets studied are large, openly branded national publishers. To respect copyright we release identifiers, hashes, code, and derived features rather than raw article text. Outlet attribution from style could be misused to flag or filter sources, and we present it as an analytic tool for studying editorial variation, not as a basis for judging the reliability or quality of any outlet.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

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 (ACL)*, pages 8440–8451.

Evan Dufraisse, Adrian Popescu, Julien Tourille, Armelle Brun, and Olivier Hamon. 2024. Combining objective and subjective perspectives for political news understanding. *arXiv preprint arXiv:2408.11174*.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Louis Escouflaire, Antonin Descampe, and Cédrick Fairon. 2024. Automated text classification of opinion vs. news french press articles. a comparison of transformer and feature-based approaches. *Language & Communication*, 99:129–140.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 503–511.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 812–818.

Shanto Iyengar. 1991. *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press, Chicago, IL.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 2479–2490, Marseille, France.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203–7219.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Masaryk University, Faculty of Informatics.

Roland Schäfer and Felix Bildhauer. 2013. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 649–657.

A Per-outlet Confusion Matrix

Figure 6 gives the row-normalized confusion matrix for the headline model (CamemBERT, temporal, entity-masked). Errors concentrate among the mutually confusable generalist Paris dailies and collapse toward Le Parisien, as summarized in Section 5.

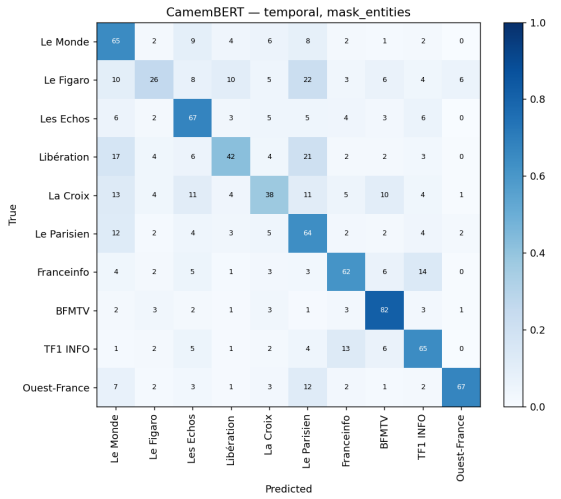


Figure 6: Row-normalized confusion matrix (CamemBERT, temporal, entity-masked). Errors concentrate among generalist Paris dailies (collapse toward Le Parisien); Ouest-France, BFMTV, Franceinfo, Les Échos retain strong diagonals.

B Fixed-model leave-one-outlet re-scoring

To assess the fingerprintability ranking under panel perturbation, we re-evaluate it under every single-outlet deletion. For each of the ten outlets we drop it from the panel, restrict the trained CamemBERT's saved class probabilities (temporal split, entity-masked, seed 42) to the nine survivors, re-argmax, and recompute per-outlet F1; no model is retrained, so the check cannot drift from the reported run. We correlate each nine-outlet ranking with the full-panel ranking of the same survivors (Table 5). Agreement is near-perfect (mean Spearman $\rho = 0.98$, min 0.93): Ouest-France or BFMTV is the most identifiable survivor in every panel and Le Figaro the most confusable whenever it is present (La Croix when Le Figaro itself is dropped). The fixed-model ranking is therefore stable under these deletions. This reuses the trained model restricted to survivors rather than retraining CamemBERT per deletion, so it probes the stability of the ranking, not small shifts in absolute accuracy.

C Tone Analysis

Figures 7 and 8 detail the tone analysis of Section 5. Tone-only features recover the outlet weakly (0.13, 1.3× chance) and add only 0.004 pooled accuracy over the surface features, and no tone dimension approaches the typographic block

Dropped	Spearman ρ	Top surv.	Bot. surv.
Le Monde	0.933	Ouest-France	Le Figaro
Le Figaro	1.000	Ouest-France	La Croix
Les Échos	1.000	Ouest-France	Le Figaro
Libération	0.983	Ouest-France	Le Figaro
La Croix	0.933	BFMTV	Le Figaro
Le Parisien	0.983	Ouest-France	Le Figaro
Franceinfo	0.983	BFMTV	Le Figaro
BFMTV	1.000	Ouest-France	Le Figaro
TF1 INFO	0.983	BFMTV	Le Figaro
Ouest-France	1.000	BFMTV	Le Figaro

Table 5: Fixed-model leave-one-outlet re-scoring of the per-outlet F1 ranking (CamemBERT, temporal, entity-masked, seed 42). ρ is the rank correlation between the nine-survivor ranking and the full-panel ranking of the same survivors. The most- and least-fingerprintable outlets are preserved throughout.

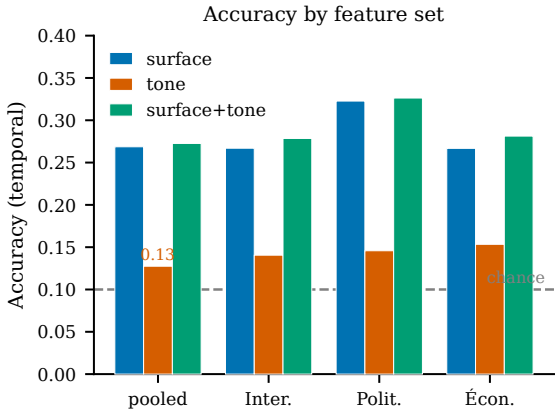


Figure 7: Tier 1 tone analysis (temporal split): accuracy by feature set and scope. Tone-only (0.13) is far weaker than surface features (0.27–0.32), and adding tone to surface lifts accuracy by only 0.004 pooled.

in weight, so the residual editorial signal is not primarily tonal under this operationalization.

D LLM Baselines: Three Families, Zero- and Few-Shot

To confirm that the zero-shot floor of Section 5 is not specific to one model family or to zero-shot prompting, we run three open-weight LLM families on the identical balanced 2025 test sample (chance = 0.100), each under matched zero-shot and few-shot prompting. Few-shot demonstrations are drawn only from the training years (< 2025), balanced across the ten outlets and three categories, and shuffled so demonstration order carries no outlet signal; the query prompt is otherwise identical to the zero-shot condition. Results are means over three seeds (42, 13, 1234); the harness,

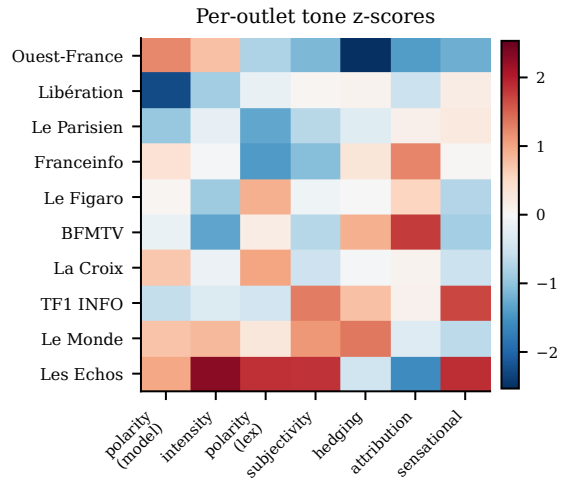


Figure 8: Per-outlet tone z-scores across seven measures (temporal split). Broadcasters (BFMTV, Franceinfo) show above-average attribution rate, but no tone dimension rivals the typographic block in discriminative weight.

prompts, and per-headline predictions are released with the code (`gen_llm_baseline.py`). This is a no-training floor, not a competitor: we therefore bound the claim to open-weight zero- and few-shot prompting. A fine-tuned or frontier model could narrow the gap to the trained encoders, but none is required to establish that outlet identity is not recoverable from a general model by prompting alone.

E Matched-Event Clustering Validation

This appendix validates the clustering pipeline (Section 4.1) and the robustness of the BFMTV agency contrast. All refits use the paper's within-event estimator and are comparable to the body's $\beta = +0.056$ ($p_{\text{Bonf}} = 0.008$). The canonical estimator and four confirmatory contrasts were prespecified (Supplementary S1); the syndication, membership, and threshold analyses here are post-hoc robustness checks, interpreted accordingly.

Cluster structure. The 626 matched events span 3,006 headlines. Clusters are small and right-skewed: a median of 4 headlines (mean 4.80, 99th percentile 9, max 12) and a median of 4 distinct outlets (mean 4.37, max 7). Because a cluster is admitted only with at least four distinct outlets, none is smaller by construction. Per-outlet coverage is uneven, from Libération (402 headlines) and La Croix (389) down to Les Échos (188) and Ouest-France (111); BFMTV contributes 290.

Model	Shots	Acc	Macro-F1	×chance
gpt-oss-120b	0-shot	.195	.189	1.9×
gpt-oss-120b	5-shot	.202	.195	2.0×
Llama-3.3-70B	0-shot	.206	.193	2.1×
Llama-3.3-70B	5-shot	.201	.188	2.0×
Mistral-Small	0-shot	.194	.171	1.9×
Mistral-Small	5-shot	.192	.163	1.9×
CamemBERT (m-ent)		.576	.569	5.8×

Table 6: Open-weight LLM outlet attribution under matched zero- and few-shot prompting, evaluated on the balanced 200-per-outlet 2025 subsample (chance exactly 0.100; mean over seeds 42, 13, 1234). Fine-tuned CamemBERT (bottom) is the trained reference. All three families sit at 1.9–2.1× chance regardless of shots; 5-shot prompting changes accuracy by at most 0.01 and never raises macro-F1, so labelled demonstrations do not help. No LLM setting approaches the trained models, consistent with the full-test zero-shot probe in Section 5: these results suggest the recoverable signal is primarily corpus-specific surface regularity rather than knowledge elicited through prompting alone.

Same-event precision and recall. On a 400-pair hand-labeled sample (184 same-event; single expert coder), same-event precision at the a priori 0.70 cutoff is 0.94 and stays ≥ 0.90 across 0.60–0.80; we keep the stricter 0.70 because looser cutoffs over-merge and dilute the within-event contrast (the threshold sweep below). Sample recall at 0.70 is 0.37, reflecting the precision-first design; clustering recall over the full 478,110-title pool remains not estimable without exhaustive same-event gold, and 30 sampled clusters were qualitatively clean (e.g. the Chinese spy-balloon and Chikirou-candidacy events). The pairs were formed by drawing random anchor headlines and pairing each with its nearest cosine neighbor, so the sample spans the near-threshold decisions the clusterer must make; a pair was coded same-event when both headlines report the same specific real-world occurrence. A single coder suffices for the bulk of this calibration, since same-event identity is a largely factual judgment and the check sets a threshold rather than reporting an outcome; even so, a second expert independently re-coded a 50-pair subset (blind to the first coder and to cosine), stratified toward the 0.60–0.80 boundary, and inter-coder agreement is high (Cohen's $\kappa = 0.82$, raw agreement 92%; 88% on the boundary band), so same-event identity is reliably codable and does not rest on one rater's judgment.

Syndication / wire copy. A same-event framing comparison is vacuous if clusters merely contain identical syndicated copy across outlets. Within each cluster we flag a headline as syndicated when it shares a word-set Jaccard ≥ 0.75 with a headline from a *different* outlet. Such cross-outlet near-duplicates are 17.6% of clustered headlines (22.8% at the looser 0.70 threshold). Removing them and re-enforcing the four-outlet minimum, the agency contrast does not weaken but *strengthens*: $\beta = +0.064$ ($p_{\text{Bonf}} = 0.023$) at Jaccard 0.75 and $\beta = +0.073$ ($p_{\text{Bonf}} = 0.016$) at 0.70, versus $+0.056$ on the full set (Table 8). The estimate remains positive after excluding near-duplicate cross-outlet copy at both tested thresholds; paraphrased or undetected shared-source text remains possible.

Cluster-membership sensitivity. Requiring ≥ 3 or ≥ 4 distinct outlets gives identical estimates ($\beta = +0.056$, $p_{\text{Bonf}} = 0.008$), since the pipeline already imposes the four-outlet floor. Tightening to ≥ 5 outlets retains only the 179 largest clusters (1,065 headlines) and the estimate attenuates to $\beta = +0.019$ (n.s.). The agency signal is thus carried by the typical four-outlet events rather than by the largest, most-covered clusters; the ≥ 5 subset is both smaller and underpowered. We read this as a scope note on where the effect concentrates, given the pre-specified ≥ 4 design, not as evidence against it.

Cosine-threshold sensitivity. Re-clustering at 0.65/0.70/0.75 yields $\beta = +0.035$ (n.s.; 741 clusters), $+0.056$ ($p_{\text{Bonf}} = 0.008$; 626), and $+0.055$ (0.016; 503): the contrast holds at and above the precision-calibrated 0.70 cutoff and dilutes only at the loosest, over-merging threshold.

Power for the binary agency estimator. The 0.12 minimum detectable effect quoted in Limitations is for the *continuous* within-event measures (valence, attribution). The surviving contrast is the *binary* `agent_subject` measure, estimated over all 3,006 clustered headlines rather than as a per-event mean, so it is more precise and warrants its own figure. Its within-event, event-clustered robust standard error is 0.019; under the pre-specified one-sided Bonferroni threshold ($\alpha = 0.05/4$) this gives a minimum detectable effect of 0.060 at 80% power and 0.068 at 90% power, roughly half the continuous-measure 0.12. The observed $\beta = 0.056$ is slightly below the nominal 0.060 80%-power MDE (attained power ≈ 0.73 at the Bonferroni level, ≈ 0.89 at the uncorrected one-sided level), so the design was powered to detect an effect of this size, not retrofit to the one that survived. The calculation is reproduced by `gen_agency_power.py` from the same within-event estimator used for the confirmatory test.

Outlet dropped	β	p_{Bonf}	clusters
(none, full panel)	$+0.056$	0.008	626
Franceinfo	$+0.029$	0.345	411
La Croix	$+0.051$	0.047	419
Le Figaro	$+0.060$	0.006	500
Le Monde	$+0.052$	0.049	423
Le Parisien	$+0.041$	0.130	416
Les Échos	$+0.059$	0.006	512
Libération	$+0.039$	0.172	377
Ouest-France	$+0.064$	0.003	554
TF1 INFO	$+0.048$	0.059	410

Table 7: Leave-one-outlet-out for the BFMTV within-event agency contrast (one-sided Bonferroni, family size 4). β stays positive in all nine drops; significance attenuates only where re-enforcing the four-outlet minimum removes the most clusters, tracking power near the $\sim$0.06 MDE. A 626-event jackknife gives $\beta \in [+0.053, +0.059]$.

Leave-one-outlet-out and event jackknife. Cluster coverage is uneven (111–402 headlines per outlet), so we check that the BFMTV agency contrast is not carried by any single outlet's coverage. Dropping each of the nine comparison outlets in turn, re-enforcing the four-outlet minimum, and refitting (Table 7), the contrast stays positive in all nine drops ($\beta \in [+0.029, +0.064]$, full panel $+0.056$) and significant in five. The four that fall short are the drops that remove the most clusters when the four-outlet minimum is re-enforced (n down to 377–416 from 626), so the attenuation tracks lost power near the estimator's $\sim$0.06 minimum detectable effect (the power note above), not a change in the effect. A leave-one-event-out jackknife (626 refits, each removing only one cluster) leaves β in $[+0.053, +0.059]$, so no single event drives it. Both checks are reproduced by `gen_agency_loo.py`.

F Is the Lexical Residual Just Beat Vocabulary?

The lexical residual of Section 5 (word choice surviving surface-stripping) could still carry outlet *topic/beat* signal rather than style: some of its

Subset	Clus.	Head.	β (p_{Bonf})
Full (≥ 4 outlets)	626	3006	$+0.056$ (0.008)
Syndic. rm., $J \geq 0.75$	434	2070	$+0.064$ (0.023)
Syndic. rm., $J \geq 0.70$	386	1832	$+0.073$ (0.016)
≥ 5 outlets	179	1065	$+0.019$ (n.s.)

Table 8: BFMTV within-event agency contrast under clustering-robustness refits. Removing cross-outlet near-duplicates (syndication) strengthens the effect; it is stable across the $\geq 3/\geq 4$ cutoff and attenuates only when restricted to the largest (≥ 5-outlet) clusters. Same estimator and pre-specified Bonferroni family as the body.

distinctive words are beat nouns (*évêques* for La Croix, *cryptos* for Les Échos) that entity masking does not remove, because they are common nouns rather than named entities. We test this directly. Starting from the surface-stripped residual (accuracy 0.324, 3.24× chance), we rank each outlet's distinctive content words by weighted log-odds with an informative Dirichlet prior (Monroe et al., 2008), estimated on the training years only, then additionally strip the top-k such words per outlet and re-fit the residual word-LR (Table 9, Figure 9).

Stripping the top-400 content words per outlet (3,413 distinct words) closes only 40% of the residual's above-chance margin; a floor of 0.234 (2.34× chance) survives. The curve flattens: doubling the words removed from 1,742 to 3,413 adds only three points of closure, indicating a persistent residual under this intervention rather than a slow decay to chance. This blocklist subsumes both the common-noun beat vocabulary of concern *and* any lowercased proper nouns entity masking may have missed (e.g. *tokyo*, *kyiv*, *seine*), so it also bounds the NER-recall concern: removing the distinctive un-masked entities does not collapse the residual. Because the log-odds blocklist additionally absorbs outlet-distinctive *register* markers (e.g. *estime* for Franceinfo, *exclusif* for Les Échos), the 40% closure cannot be attributed solely to beat vocabulary. Substantial lexical signal remains after the tested blocklist, but may still encode beat through unstripped words or correlated lexical structure; the framing component is isolated separately by the matched-event design (Section 5.4), which does not rely on this residual.

top-k/outlet	words	acc	×chance	margin closed
0	0	.324	3.24×	0%
25	228	.285	2.85×	17%
50	442	.271	2.71×	24%
100	873	.258	2.58×	29%
200	1,742	.241	2.41×	37%
400	3,413	.234	2.34×	40%

Table 9: Residual accuracy as the top-k outlet-distinctive content words per outlet are stripped (word-LR, temporal split, chance = 0.100). Even after removing 3,413 distinct words, a 2.34×-chance residual survives; the tested blocklist closes 40% of the above-chance margin and also removes some register markers.

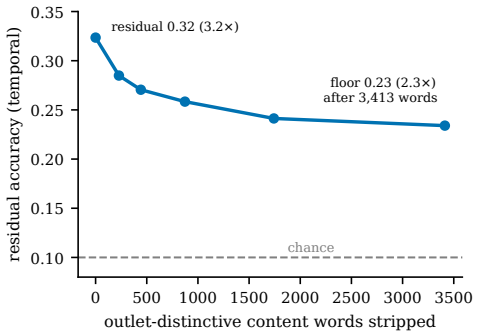


Figure 9: Residual accuracy versus number of outlet-distinctive content words stripped. The curve flattens toward 2.34× chance under the tested blocklist intervention.

A Taxonomy of Programming Languages for Code Generation

Nishat Raihan[1], **Christian Newman**[2], **Marcos Zampieri**[1]
[1]George Mason University, USA
[2]Rochester Institute of Technology, USA
mraihan2@gmu.edu

Abstract

The world's 7,000+ languages vary widely in the availability of resources for NLP, motivating efforts to systematically categorize them by their degree of resourcefulness (Joshi et al., 2020). A similar disparity exists among programming languages (PLs); however, no resource-tier taxonomy has been established for code. As large language models (LLMs) grow increasingly capable of generating code, such a taxonomy becomes essential. To fill this gap, we present the first reproducible PL resource classification, grouping 646 languages into four tiers. We show that only 1.9% of languages (Tier 3, High) account for 74.6% of all tokens in seven major web-derived corpora, while 71.7% of languages (Tier 0, Scarce) contribute just 1.0%. Statistical analyses of within-tier inequality, dispersion, and distributional skew confirm that this imbalance is both extreme and systematic. Our results provide a principled framework for dataset curation and tier-aware evaluation of multilingual LLMs, and a first quantitative picture of code availability on the web.

1 Introduction

The advent of large language models trained on code (Code LLMs) has triggered a paradigm shift in software engineering (Chen et al., 2021; Li et al., 2022; Raihan et al., 2024b). These models, powered by vast web-scraped corpora such as The Stack (Kocetkov et al., 2022) and CodeSearchNet (Husain et al., 2019), can generate complex code, translate between programming languages, and accelerate software development. However, their capabilities are contingent on their training data, which is heavily skewed toward a few popular PLs like Python and JavaScript. This imbalance, noted in analyses of code corpora (Madaan et al., 2022), creates a "digital divide" where resource-rich languages receive superior model support while less common languages lag behind, with potential downstream implications for both fairness and security (Fan et al., 2023; Pearce et al., 2025).

While the existence of this disparity is acknowledged, the field lacks a systematic framework to quantify it. This challenge mirrors the well-documented problem of resource scarcity in NLP. The work by Joshi et al. (2020) provides a data-driven classification of natural languages into resource tiers, becoming a foundational tool for measuring linguistic diversity and guiding efforts to support low-resource languages. We argue that the PL ecosystem requires a similar foundational effort to move from anecdotal observations to quantitative understanding, enabling rigorous evaluation of models on less-represented languages (Zhao and Fard, 2025) and fostering development of truly multilingual code models (Nijkamp et al., 2023).

To address this gap, we present the first large-scale classification of PLs based on their prevalence in the web-derived corpora that power modern Code LLMs. We collect, merge, and deduplicate seven prominent code datasets and quantify each language's presence via token counts, a direct measure of availability on the web. From this analysis, we introduce a four-tier classification scheme: *high-*, *medium-*, *low-*, and *scarce-resource*. This work makes three primary contributions:

C1 A comprehensive empirical analysis of PL distribution across modern code corpora.

C2 A clear, data-driven four-tier classification of resource availability for 646 PLs.

C3 A public release of our token counts and tier assignments to facilitate equitable model benchmarking, bias diagnosis, and targeted research into the long tail of PLs.

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 37–52
October 29, 2026 ©2026 Association for Computational Linguistics

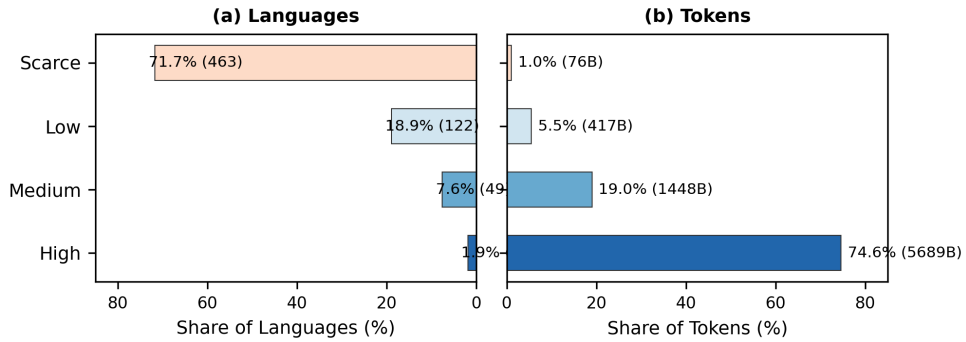


Figure 1: Resource imbalance across four tiers. **(a)**: share of languages; **(b)**: share of tokens. Only 1.9% of languages (High) supply 74.6% of all tokens, while 71.7% of languages (Scarce) contribute just 1.0%.

2 Related Work

Taxonomies for Natural Languages The foundational work of Joshi et al. (2020) introduced a six-class taxonomy for natural languages based on the availability of labeled and unlabeled data. Their framework has since guided data collection, model evaluation, and resource allocation across low-resource NLP. Our work adapts this philosophy to the PL domain, where data imbalance is equally pronounced but has lacked formal characterization.

Code Generation Data The rapid scaling of code pre-training data, from CodeSearchNet's six languages (Husain et al., 2019) to The Stack v2's 619 (Lozhkov et al., 2024), has enabled increasingly capable Code LLMs (Chen et al., 2021; Li et al., 2023; Roziere et al., 2023; Guo et al., 2024). However, the language composition of these corpora remains opaque. Madaan et al. (2022) and Raihan et al. (2024b) note that evaluations are overwhelmingly Python-centric, and recent surveys confirm that nearly 70% of Code LLM studies involve Python-generated code (Raihan et al., 2025). Our taxonomy addresses this gap by providing a quantitative inventory of language representation across seven major corpora.

Multilingual Code Evaluation Benchmarks such as MultiPL-E (Cassano et al., 2023), HumanEval-X (Zheng et al., 2023), mHumanEval Raihan et al. (2024a), and the multilingual evaluation of Athiwaratkun et al. (2023) have extended code evaluation beyond Python. While these efforts are valuable, they remain limited to a small subset of "popular" PLs. Our taxonomy provides the missing link: a principled basis for selecting which languages to include and for interpreting performance differences through the lens of training-data availability.

3 Methods

3.1 Data Collection

Our analysis draws on seven corpora selected to represent the modern Code LLM data ecosystem (Table 1). The collection spans massive multilingual pre-training sets like The Stack v2 (Lozhkov et al., 2024) and RefineCode (Huang et al., 2025), each covering over 600 languages with diverging curation philosophies: one based on exhaustive, graph-based collection and the other on aggressive filtering for safety and reproducibility.

Corpora with lower language counts serve specialized roles. StarCoder Data (Li et al., 2023) provides a curated 86-language subset. Project CodeNet (Puri et al., 2021) offers rich problem-centric metadata for reasoning tasks, while CodeSearchNet (Husain et al., 2019) remains vital for code-retrieval models. The Heap (Katzy et al., 2025) is indispensable not as a training source but as a contamination-free evaluation benchmark. GitHub Code (CodeParrot) (Touvron et al., 2023) provides a direct snapshot of public repository content. This diversity in scale, purpose, and curation philosophy provides a robust foundation for our classification.

3.2 Corpus Consolidation and Token Accounting

To obtain reliable statistics, we fuse the seven raw corpora into a single cleaned dataset and measure it with a code-aware tokenizer. The workflow proceeds in three stages:

S1 Merge and Deduplicate. We combine all corpora and drop exact duplicates by comparing SHA-256 hashes at file level, ensuring each unique source file contributes exactly once.

S2 PL Tagging. We infer the programming language from the file extension using a refined

Corpus	Reported Size	# of PLs	Reference
The Stack v2	900 B tokens	619	Lozhkov et al. (2024)
RefineCode	960 B tokens	607	Huang et al. (2025)
StarCoder Data	250 B tokens, 783 GB	86	Li et al. (2023)
The Heap	32.7 M files	57	Katzy et al. (2025)
Project CodeNet	14 M samples, 500 M LOC	55	Puri et al. (2021)
GitHub Code (CodeParrot)	1 TB, 115 M files	32	Touvron et al. (2023)
CodeSearchNet Corpus	20 GB, 6 M functions	6	Husain et al. (2019)

Table 1: Overview of the seven code corpora used in our study, sorted by number of programming languages (PLs). The selection represents the diversity of sources for training and evaluating modern Code LLMs.

map based on GitHub Linguist, which scales to millions of files while maintaining high precision across 600+ languages.

S3 Token Accounting. We tokenize every file with the StarCoder tokenizer (Li et al., 2023), record per-language token counts, and log auxiliary metrics (files, lines, bytes) for quality checks. Because the tokenizer mirrors how a competitive Code LLM "sees" code, these counts serve as a faithful proxy for model exposure.

4 Classification and Distribution

For every one of the 646 programming languages L in our corpus, we measure the total number of *source-code tokens*, t_L. Adapting the resource taxonomy for natural languages (Joshi et al., 2020), we divide languages into four tiers:

$$\begin{aligned}
&\text{Tier}_3 : t_L \geq 100\text{B};\\
&\text{Tier}_2 : 10\text{B} \leq t_L < 100\text{B};\\
&\text{Tier}_1 : 1\text{B} \leq t_L < 10\text{B};\\
&\text{Tier}_0 : t_L < 1\text{B}.
\end{aligned}$$

The corpus shows a pronounced long tail (Figure 1): Tier 0 covers 71.7% of the 646 languages yet supplies only 1.0% of the 7.63T tokens, whereas the 12 Tier 3 languages (1.9%) supply 74.6%. Table 2 summarizes the four tiers of taxonomy.

Tier	# PLs	Tokens (B)	Notable Examples
High	12	5,689	Python, JS, Java, C#, C++
Medium	49	1,448	SQL, Go, Kotlin, Rust, R
Low	122	417	Prolog, Fortran, Julia, OCaml
Scarce	463	76	Ada, COBOL, Awk, GAMS

Table 2: Four-tier resource classification. Only 1.9% of languages (High) account for 74.6% of all 7.63T tokens; 71.7% of languages (Scarce) contribute just 1.0%.

Table 3 lists the ten most-resourced languages and the ten least-resourced languages studied. The top-10 languages together hold 71.6% of all tokens. Benchmarks dominated by these languages risk overstating model robustness (Cassano et al., 2023; Zheng et al., 2023; Athiwaratkun et al., 2023). On the other hand, the ten least-resourced languages together account for only 0.00003% of tokens (roughly 2.02M of 7.63T). Despite their rarity, these niche formats and DSLs (e.g., templates, tracebacks, meta-languages) are vital for real-world tooling and software analysis and merit targeted data collection. The full taxonomy is presented in Appendix A.

Rank	Language	Tier	Resource
1	Python	3	High
2	JavaScript	3	High
3	Java	3	High
4	C#	3	High
5	PHP	3	High
6	C++	3	High
7	CSS	3	High
8	TypeScript	3	High
9	C	3	High
10	Ruby	3	High
637	Myghty	0	Scarce
638	Object Data Instance Notation	0	Scarce
639	Ecere Projects	0	Scarce
640	NumPy	0	Scarce
641	Omgrofl	0	Scarce
642	Python traceback	0	Scarce
643	MiniD	0	Scarce
644	NetLinx+ERB	0	Scarce
645	Record Jar	0	Scarce
646	C-ObjDump	0	Scarce

Table 3: Top and bottom-10 programming languages.

5 Significance and Analysis

State-of-the-art Code LLMs, such as StarCoder (Li et al., 2023) and Code Llama (Roziere et al., 2023), train on multi-terabyte corpora whose language mix is opaque and heavily skewed toward a handful of "major" PLs. Inspired by the NLP resource taxonomy of Joshi et al. (2020), we introduce a

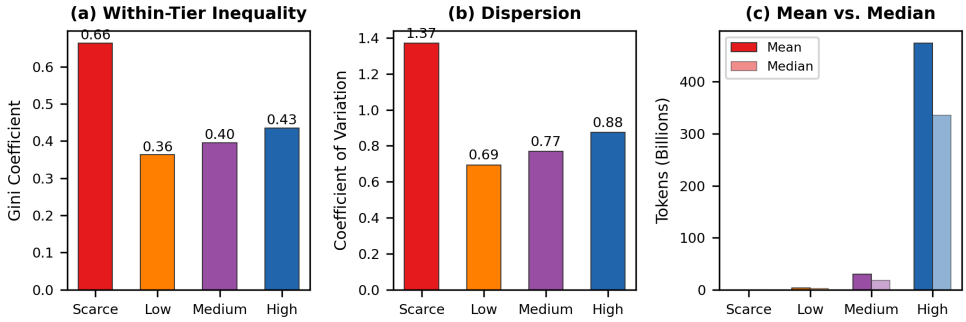


Figure 2: Within-tier statistics: **(a)** Gini coefficient measuring token inequality; **(b)** coefficient of variation measuring dispersion; **(c)** mean vs. median tokens per language, highlighting right-skew in every tier.

transparent, token-count-based classification that *systematically* exposes and quantifies this skew, rather than relying on ad hoc labels.

Imbalances The Gini coefficient (Figure 2a) reveals that inequality is most severe within the Scarce tier (Gini = 0.66), where a few languages near the 1B boundary dwarf hundreds of near-zero entries. The High tier also shows concentration (Gini = 0.43), since Python alone holds ∼27% of Tier 3 tokens. Without such diagnostics, aggregate metrics would be dominated by Python and C++ while dozens of low-resource PLs remain invisible (Muennighoff et al., 2023; Raihan et al., 2024a).

Variation The coefficient of variation (Figure 2b) identifies which tiers contain the most heterogeneous PL populations. The Scarce tier's CV of 1.37 signals that targeted scraping, augmentation, or synthetic generation could disproportionately benefit the most under-served languages. The mean-vs.-median comparison (Figure 2c) further highlights the right-skewed nature of every tier: means consistently exceed medians, confirming that a few dominant languages inflate tier-level averages.

Tier-aware Evaluation. The Lorenz curves (Figure 3) show that deviation from perfect equality is most pronounced in the High and Medium tiers, indicating that even within resource-rich groupings, token allocation is far from uniform. The ECDF survival curves (Figure 4) complement this view: at a threshold of 10B tokens, 100% of High-tier languages survive but only ∼45% of Medium-tier languages do, with Low and Scarce tiers dropping sharply below 1B. These visualizations motivate stratified scoring in benchmarks and leaderboards, rather than relying on overall means that mask per-tier weaknesses.

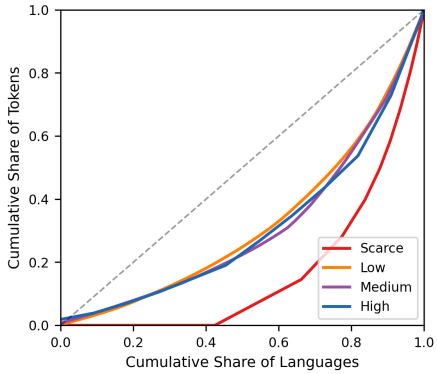


Figure 3: Curves of token distribution within each tier; deviation from the diagonal indicates concentration.

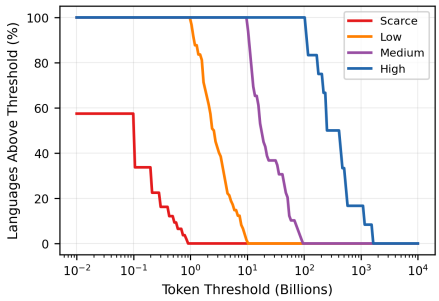


Figure 4: Percentage of languages above token thresholds (log scale), comparing survivability across tiers.

Infrastructure As corpora like The Stack v2 (Lozhkov et al., 2024) continue to grow, the taxonomy offers a reproducible scaffold to diagnose bias and track progress. It also provides a lens for interpreting model behavior: when a Code LLM defaults to Python under a language-neutral prompt, the taxonomy quantifies *why*, since Python commands 1,543B tokens (20.2% of the corpus), dwarfing all other languages. By linking training-data composition to generation bias, the taxonomy transforms "Python default" from an anecdote into a measurable, tier-level phenomenon.

6 Conclusion

We present the first tiered, token-count-based taxonomy to categorize PL resource levels. Our analysis of seven major web-derived code corpora covering 646 PLs and 7.63 trillion tokens reveals stark imbalance: 12 High-resource languages hold 74.6% of all tokens, while 463 Scarce-resource languages share just 1.0%. Gini coefficients, Lorenz curves, and CV analysis confirm that this imbalance persists both across and within tiers.

Existing multilingual benchmarks (Cassano et al., 2023; Zheng et al., 2023; Athiwaratkun et al., 2023; Raihan et al., 2024a) evaluate beyond Python but select languages without accounting for training-data availability. Our taxonomy is the first framework to not only identify under-resourced PLs but quantify the magnitude of that scarcity. Viewed through a web-as-corpus lens, it also documents which PLs the web over- and under-represents, guiding future crawling and curation efforts.

Limitations

Our choices of seven permissively licensed corpora, a single StarCoder tokenizer, extension-based PL detection, exact-hash deduplication, and heuristic tier cutoffs are intentional for transparency, reproducibility, and alignment with how current Code LLMs are actually trained. These design points may influence absolute counts (e.g., polyglot files, boilerplate variants), but they do not disturb the heavy-tailed shape or the rank-order skew our taxonomy exposes. We release raw counts so anyone can swap tokenizers, tweak thresholds, or apply finer-grained deduplication without rerunning the entire pipeline. In short, these are controlled methodological knobs: they might refine decimal values, but they do not overturn the central finding that code resources are starkly imbalanced and demand systematic accounting.

Ethical Considerations

We follow the ACL Code of Ethics and use only permissively licensed, publicly accessible data, respecting original licenses and platform terms of service. No proprietary or personal data were intentionally included, and we release only aggregate statistics, not raw code, to minimize privacy and licensing risks.

References

Ben Athiwaratkun, Sanjay Krishna Gouda, Zijian Wang, Xiaopeng Li, Yuchen Tian, et al. 2023. Multi-lingual evaluation of code generation models. *Proceedings of ICLR*.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, et al. 2023. Multipl-e: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49:3675–3691.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang. 2023. Large language models for software engineering: Survey and open problems. In *Proceedings of ICSE*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, et al. 2024. Deepseek-coder: when the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Siming Huang, Tianhao Cheng, Jason Klein Liu, Weidi Xu, et al. 2025. Opencoder: The open cookbook for top-tier code large language models. In *Proceedings of ACL*.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of ACL*.

Jonathan Katzy, Razvan Mihai Popescu, Arie van Deursen, and Maliheh Izadi. 2025. The heap: A contamination-free multilingual code dataset for evaluating large language models. In *Proceedings of FORGE*.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, et al. 2022. The stack: 3 TB of permissively licensed source code. *TMLR*.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, et al. 2023. StarCoder: May the source be with you! *TMLR*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, et al. 2022. Competition-level code generation with alphacode. *Science*, 378:1092–1097.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. In *Proceedings of EMNLP*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of ACL*.

Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309*.

Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2025. Asleep at the keyboard? assessing the security of github copilot's code contributions. *Communications of the ACM*.

Ruchir Puri, David S Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, et al. 2021. Codenet: A large-scale ai for code dataset for learning a diversity of coding tasks. In *Proceedings of NeurIPS*.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2024a. mHumanEval: A multilingual benchmark for code generation. In *Findings of the ACL (NAACL-2025)*.

Nishat Raihan, Christian Newman, and Marcos Zampieri. 2024b. Code llms: A taxonomy-based survey. In *Proceedings of BigData*.

Nishat Raihan, Mohammed Latif Siddiq, Joanna CS Santos, and Marcos Zampieri. 2025. Large language models in computer science education: A systematic literature review. In *Proceedings of SIGCSE*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zixiao Zhao and Fatemeh Fard. 2025. Do current language models support code intelligence for r programming language? *ACM Transactions on Software Engineering and Methodology*, 34.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zihan Wang, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of KDD*.

A The Full PL Classification

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
1	Python	154281537	1542.8	3	High
2	JavaScript	108871932	1088.7	3	High
3	Java	56933829	569.3	3	High
4	C#	51233056	512.3	3	High
5	PHP	46025156	460.3	3	High
6	C++	43175936	431.8	3	High
7	CSS	23869063	238.7	3	High
8	TypeScript	23853979	238.5	3	High
9	C	20775502	207.8	3	High
10	Ruby	17785998	177.9	3	High
11	Objective-C	11407073	114.1	3	High
12	Shell	10681876	106.8	3	High
13	Go	9301819	93.0	2	Medium
14	Kotlin	8616859	86.2	2	Medium
15	Jupyter Notebook	8151402	81.5	2	Medium
16	Vue	7375658	73.8	2	Medium
17	Swift	7234581	72.3	2	Medium
18	TSX	5496011	55.0	2	Medium
19	Java Server Pages	5394783	53.9	2	Medium
20	SQL	5285705	52.9	2	Medium
21	INI	5261835	52.6	2	Medium
22	R	5145606	51.5	2	Medium
23	Blade	4925151	49.3	2	Medium
24	MATLAB	4543828	45.4	2	Medium
25	Scala	4445036	44.5	2	Medium
26	Makefile	4221531	42.2	2	Medium
27	Dart	4214307	42.1	2	Medium
28	Java Properties	3611528	36.1	2	Medium
29	HTML+Razor	3510527	35.1	2	Medium
30	TeX	3190550	31.9	2	Medium
31	Lua	2347887	23.5	2	Medium
32	Rust	2218033	22.2	2	Medium
33	TSV	2116969	21.2	2	Medium
34	Smali	2039732	20.4	2	Medium
35	Dockerfile	1898900	19.0	2	Medium
36	reStructuredText	1826351	18.3	2	Medium
37	Hack	1786841	17.9	2	Medium
38	CMake	1743194	17.4	2	Medium
39	Less	1666356	16.7	2	Medium
40	Microsoft Visual Studio Solution	1665208	16.7	2	Medium
41	WebVTT	1603735	16.0	2	Medium
42	Gradle	1598251	16.0	2	Medium
43	Pug	1503380	15.0	2	Medium
44	EJS	1497213	15.0	2	Medium
45	Twig	1316871	13.2	2	Medium
46	ASP.NET	1287527	12.9	2	Medium
47	Haskell	1254071	12.5	2	Medium
48	Diff	1204035	12.0	2	Medium
49	Gherkin	1203914	12.0	2	Medium
50	Handlebars	1188119	11.9	2	Medium
51	TSQL	1155064	11.6	2	Medium
52	Wikitext	1147269	11.5	2	Medium
53	Perl	1146878	11.5	2	Medium
54	Batchfile	1127069	11.3	2	Medium
55	Groovy	1103831	11.0	2	Medium
56	Public Key	1062027	10.6	2	Medium
57	Visual Basic .NET	1060179	10.6	2	Medium
58	Haml	1046278	10.5	2	Medium
59	CoffeeScript	1027516	10.3	2	Medium
60	Smalltalk	1025689	10.3	2	Medium
61	ActionScript	1013150	10.1	2	Medium
62	RMarkdown	994233	9.9	1	Low
63	Wavefront Object	984476	9.8	1	Low

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
64	Pascal	951936	9.5	1	Low
65	PostScript	950368	9.5	1	Low
66	Roff	936321	9.4	1	Low
67	Org	893315	8.9	1	Low
68	Gerber Image	887319	8.9	1	Low
69	Elixir	881706	8.8	1	Low
70	Checksums	833913	8.3	1	Low
71	TOML	815905	8.2	1	Low
72	Assembly	773883	7.7	1	Low
73	Smarty	769939	7.7	1	Low
74	Sass	756458	7.6	1	Low
75	Maven POM	749908	7.5	1	Low
76	HCL	749763	7.5	1	Low
77	PowerShell	682789	6.8	1	Low
78	Pickle	643341	6.4	1	Low
79	Verilog	636016	6.4	1	Low
80	Processing	578171	5.8	1	Low
81	Vim Script	572386	5.7	1	Low
82	mIRC Script	561872	5.6	1	Low
83	GLSL	548969	5.5	1	Low
84	Roff Manpage	525533	5.3	1	Low
85	Game Maker Language	500777	5.0	1	Low
86	Unix Assembly	494749	4.9	1	Low
87	Gettext Catalog	490963	4.9	1	Low
88	HTML+PHP	478802	4.8	1	Low
89	VHDL	451684	4.5	1	Low
90	Julia	447047	4.5	1	Low
91	Motorola 68K Assembly	442526	4.4	1	Low
92	Graphviz (DOT)	433493	4.3	1	Low
93	Scheme	426836	4.3	1	Low
94	Apex	408075	4.1	1	Low
95	OpenEdge ABL	403912	4.0	1	Low
96	Go Module	398105	4.0	1	Low
97	AsciiDoc	396165	4.0	1	Low
98	Stylus	395048	4.0	1	Low
99	CoNLL-U	393103	3.9	1	Low
100	Clojure	389392	3.9	1	Low
101	Slim	367031	3.7	1	Low
102	Haxe	366379	3.7	1	Low
103	Emacs Lisp	365757	3.7	1	Low
104	OCaml	364452	3.6	1	Low
105	SubRip Text	362904	3.6	1	Low
106	Git Config	360341	3.6	1	Low
107	Fortran Free Form	352647	3.5	1	Low
108	AMPL	352293	3.5	1	Low
109	Jest Snapshot	318704	3.2	1	Low
110	Common Lisp	315454	3.2	1	Low
111	BibTeX	314691	3.1	1	Low
112	Rich Text Format	305866	3.1	1	Low
113	Erlang	303085	3.0	1	Low
114	GDScript	284996	2.8	1	Low
115	Jinja	278431	2.8	1	Low
116	JSON with Comments	277494	2.8	1	Low
117	Fortran	271017	2.7	1	Low
118	F#	266595	2.7	1	Low
119	Fluent	264102	2.6	1	Low
120	Puppet	255916	2.6	1	Low
121	XML Property List	255867	2.6	1	Low
122	JAR Manifest	255357	2.6	1	Low
123	D	253920	2.5	1	Low
124	VBA	244935	2.4	1	Low
125	edn	240030	2.4	1	Low
126	desktop	239371	2.4	1	Low
127	Svelte	234670	2.3	1	Low
128	Gnuplot	234425	2.3	1	Low
129	Protocol Buffer	234297	2.3	1	Low

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
130	PLSQL	231741	2.3	1	Low
131	Nix	230308	2.3	1	Low
132	M	223873	2.2	1	Low
133	PureBasic	220170	2.2	1	Low
134	Turtle	213057	2.1	1	Low
135	Raku	211082	2.1	1	Low
136	Raw token data	207118	2.1	1	Low
137	Tcl	204437	2.0	1	Low
138	mcfunction	201109	2.0	1	Low
139	JavaScript+ERB	198008	2.0	1	Low
140	Racket	197694	2.0	1	Low
141	Classic ASP	193001	1.9	1	Low
142	Prolog	187485	1.9	1	Low
143	QMake	187415	1.9	1	Low
144	Gentoo Ebuild	185598	1.9	1	Low
145	PLpgSQL	183383	1.8	1	Low
146	QML	182471	1.8	1	Low
147	Solidity	180287	1.8	1	Low
148	DIGITAL Command Language	179931	1.8	1	Low
149	SQF	173685	1.7	1	Low
150	Elm	173109	1.7	1	Low
151	SystemVerilog	171959	1.7	1	Low
152	Cuda	171244	1.7	1	Low
153	Procfile	170824	1.7	1	Low
154	KiCad Layout	167029	1.7	1	Low
155	Mathematica	166560	1.7	1	Low
156	RPM Spec	165787	1.7	1	Low
157	XSLT	165580	1.7	1	Low
158	Standard ML	164588	1.6	1	Low
159	GraphQL	160523	1.6	1	Low
160	Groovy Server Pages	157685	1.6	1	Low
161	E-mail	149672	1.5	1	Low
162	ApacheConf	148905	1.5	1	Low
163	COLLADA	146726	1.5	1	Low
164	Starlark	134350	1.3	1	Low
165	Objective-C++	133114	1.3	1	Low
166	Scilab	132876	1.3	1	Low
167	Liquid	130642	1.3	1	Low
168	PlantUML	129291	1.3	1	Low
169	IDL	122615	1.2	1	Low
170	Wavefront Material	121616	1.2	1	Low
171	Stata	121052	1.2	1	Low
172	STL	119027	1.2	1	Low
173	fish	116039	1.2	1	Low
174	LLVM	113925	1.1	1	Low
175	HLSL	113570	1.1	1	Low
176	xBase	113188	1.1	1	Low
177	Forth	109194	1.1	1	Low
178	Mustache	109049	1.1	1	Low
179	ColdFusion	105315	1.1	1	Low
180	FreeMarker	104830	1.0	1	Low
181	Crystal	102607	1.0	1	Low
182	Nunjucks	100455	1.0	1	Low
183	Git Attributes	97737	1.0	1	Low
184	AppleScript	94262	0.9	0	Scarce
185	1C Enterprise	94066	0.9	0	Scarce
186	Coq	92375	0.9	0	Scarce
187	YASnippet	92294	0.9	0	Scarce
188	Ada	92104	0.9	0	Scarce
189	X PixMap	90219	0.9	0	Scarce
190	Web Ontology Language	89435	0.9	0	Scarce
191	LilyPond	87374	0.9	0	Scarce
192	robots.txt	85086	0.9	0	Scarce
193	SaltStack	82776	0.8	0	Scarce
194	Ant Build System	82227	0.8	0	Scarce
195	Nim	81595	0.8	0	Scarce

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
196	ReScript	80682	0.8	0	Scarce
197	UnrealScript	79600	0.8	0	Scarce
198	RDoc	79514	0.8	0	Scarce
199	ShaderLab	77504	0.8	0	Scarce
200	KiCad Schematic	75638	0.8	0	Scarce
201	OpenSCAD	74708	0.7	0	Scarce
202	HAProxy	74344	0.7	0	Scarce
203	RobotFramework	73926	0.7	0	Scarce
204	Go Checksums	71195	0.7	0	Scarce
205	AGS Script	71041	0.7	0	Scarce
206	SAS	69655	0.7	0	Scarce
207	Pure Data	69308	0.7	0	Scarce
208	Yacc	68492	0.7	0	Scarce
209	AngelScript	67669	0.7	0	Scarce
210	Gemfile.lock	67462	0.7	0	Scarce
211	Tcsh	66589	0.7	0	Scarce
212	Vim Snippet	65973	0.7	0	Scarce
213	HTTP	65392	0.7	0	Scarce
214	Nginx	64585	0.6	0	Scarce
215	Textile	64490	0.6	0	Scarce
216	Cabal Config	64393	0.6	0	Scarce
217	AutoHotkey	64004	0.6	0	Scarce
218	GAP	63667	0.6	0	Scarce
219	SMT	63299	0.6	0	Scarce
220	HTML+EEX	63121	0.6	0	Scarce
221	BitBake	61541	0.6	0	Scarce
222	Cython	60384	0.6	0	Scarce
223	PureScript	60146	0.6	0	Scarce
224	Lex	59656	0.6	0	Scarce
225	Max	59510	0.6	0	Scarce
226	Reason	57730	0.6	0	Scarce
227	G-code	54523	0.5	0	Scarce
228	VBScript	53900	0.5	0	Scarce
229	NWScript	52938	0.5	0	Scarce
230	ColdFusion CFC	49873	0.5	0	Scarce
231	Xtend	49869	0.5	0	Scarce
232	Modelica	49219	0.5	0	Scarce
233	Eagle	49207	0.5	0	Scarce
234	Papyrus	48386	0.5	0	Scarce
235	ABAP	46774	0.5	0	Scarce
236	nesC	46665	0.5	0	Scarce
237	SWIG	46645	0.5	0	Scarce
238	JSONLD	45703	0.5	0	Scarce
239	SuperCollider	45464	0.5	0	Scarce
240	Rebol	44687	0.4	0	Scarce
241	Agda	43197	0.4	0	Scarce
242	EditorConfig	42796	0.4	0	Scarce
243	XQuery	42556	0.4	0	Scarce
244	SourcePawn	42411	0.4	0	Scarce
245	CSON	40598	0.4	0	Scarce
246	SELinux Policy	40594	0.4	0	Scarce
247	Linux Kernel Module	40477	0.4	0	Scarce
248	Pawn	40283	0.4	0	Scarce
249	AL	40073	0.4	0	Scarce
250	Logos	40071	0.4	0	Scarce
251	AspectJ	39898	0.4	0	Scarce
252	Awk	39800	0.4	0	Scarce
253	Protocol Buffer Text Format	39461	0.4	0	Scarce
254	AutoIt	39112	0.4	0	Scarce
255	Vim Help File	38710	0.4	0	Scarce
256	OpenCL	38697	0.4	0	Scarce
257	LiveScript	36575	0.4	0	Scarce
258	Windows Registry Entries	36440	0.4	0	Scarce
259	COBOL	34905	0.3	0	Scarce
260	M4Sugar	34688	0.3	0	Scarce
261	Eiffel	34604	0.3	0	Scarce

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
262	Linker Script	34355	0.3	0	Scarce
263	BASIC	34168	0.3	0	Scarce
264	LookML	33854	0.3	0	Scarce
265	Latte	33347	0.3	0	Scarce
266	Cue Sheet	32755	0.3	0	Scarce
267	DirectX 3D File	32664	0.3	0	Scarce
268	JetBrains MPS	31818	0.3	0	Scarce
269	RAML	31284	0.3	0	Scarce
270	Volt	30361	0.3	0	Scarce
271	Graph Modeling Language	30282	0.3	0	Scarce
272	Ren'Py	30228	0.3	0	Scarce
273	SPARQL	29134	0.3	0	Scarce
274	HiveQL	28582	0.3	0	Scarce
275	AIDL	28577	0.3	0	Scarce
276	Lean	28334	0.3	0	Scarce
277	Common Workflow Language	27821	0.3	0	Scarce
278	DM	27296	0.3	0	Scarce
279	World of Warcraft Addon Data	27220	0.3	0	Scarce
280	GN	27184	0.3	0	Scarce
281	PostCSS	27016	0.3	0	Scarce
282	DTrace	27009	0.3	0	Scarce
283	kvlang	26345	0.3	0	Scarce
284	ZenScript	26187	0.3	0	Scarce
285	FreeBasic	25757	0.3	0	Scarce
286	MAXScript	25610	0.3	0	Scarce
287	CODEOWNERS	25378	0.3	0	Scarce
288	Alpine Abuild	24652	0.2	0	Scarce
289	Gosu	24554	0.2	0	Scarce
290	Portugol	24373	0.2	0	Scarce
291	PowerBuilder	24296	0.2	0	Scarce
292	ANTLR	24102	0.2	0	Scarce
293	Mako	24004	0.2	0	Scarce
294	POV-Ray SDL	23883	0.2	0	Scarce
295	X BitMap	23690	0.2	0	Scarce
296	Glyph Bitmap Distribution Format	23321	0.2	0	Scarce
297	ObjectScript	23153	0.2	0	Scarce
298	CWeb	22893	0.2	0	Scarce
299	Dhall	22195	0.2	0	Scarce
300	Idris	22068	0.2	0	Scarce
301	NewLisp	21694	0.2	0	Scarce
302	Meson	21524	0.2	0	Scarce
303	Slash	21422	0.2	0	Scarce
304	Ballerina	21341	0.2	0	Scarce
305	Turing	21270	0.2	0	Scarce
306	LabVIEW	21172	0.2	0	Scarce
307	Zig	20797	0.2	0	Scarce
308	sed	20692	0.2	0	Scarce
309	M4	20666	0.2	0	Scarce
310	Objective-J	19987	0.2	0	Scarce
311	Pod	19974	0.2	0	Scarce
312	Csound Document	19932	0.2	0	Scarce
313	CodeQL	19180	0.2	0	Scarce
314	F*	19084	0.2	0	Scarce
315	q	19023	0.2	0	Scarce
316	4D	18929	0.2	0	Scarce
317	GSC	18087	0.2	0	Scarce
318	ASL	18058	0.2	0	Scarce
319	Kit	17739	0.2	0	Scarce
320	Jasmin	17590	0.2	0	Scarce
321	Proguard	17548	0.2	0	Scarce
322	Literate Haskell	17195	0.2	0	Scarce
323	RPGLE	17136	0.2	0	Scarce
324	Adobe Font Metrics	17030	0.2	0	Scarce
325	YANG	17016	0.2	0	Scarce
326	Stan	16966	0.2	0	Scarce
327	NEON	16959	0.2	0	Scarce

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
328	wdl	16876	0.2	0	Scarce
329	NCL	16626	0.2	0	Scarce
330	V	16619	0.2	0	Scarce
331	CLIPS	16339	0.2	0	Scarce
332	Isabelle	16299	0.2	0	Scarce
333	Euphoria	15967	0.2	0	Scarce
334	Thrift	15912	0.2	0	Scarce
335	Squirrel	15866	0.2	0	Scarce
336	OpenQASM	15673	0.2	0	Scarce
337	HXML	15594	0.2	0	Scarce
338	Io	15147	0.2	0	Scarce
339	Vala	15079	0.2	0	Scarce
340	OpenStep Property List	14926	0.1	0	Scarce
341	Velocity Template Language	14882	0.1	0	Scarce
342	Nextflow	14831	0.1	0	Scarce
343	PigLatin	14665	0.1	0	Scarce
344	Jsonnet	14335	0.1	0	Scarce
345	Clarion	14109	0.1	0	Scarce
346	NPM Config	13594	0.1	0	Scarce
347	Astro	13465	0.1	0	Scarce
348	MoonScript	13417	0.1	0	Scarce
349	Limbo	13261	0.1	0	Scarce
350	Cool	13253	0.1	0	Scarce
351	Factor	13160	0.1	0	Scarce
352	Lasso	12983	0.1	0	Scarce
353	Clean	12920	0.1	0	Scarce
354	ChucK	12882	0.1	0	Scarce
355	Redirect Rules	12856	0.1	0	Scarce
356	Pan	12821	0.1	0	Scarce
357	DNS Zone	12790	0.1	0	Scarce
358	Arc	12296	0.1	0	Scarce
359	MLIR	12290	0.1	0	Scarce
360	WebAssembly	12278	0.1	0	Scarce
361	Mercury	11898	0.1	0	Scarce
362	YARA	11810	0.1	0	Scarce
363	Inno Setup	11765	0.1	0	Scarce
364	MQL4	11585	0.1	0	Scarce
365	OpenType Feature File	11411	0.1	0	Scarce
366	Promela	11300	0.1	0	Scarce
367	Marko	11114	0.1	0	Scarce
368	Wollok	11078	0.1	0	Scarce
369	MQL5	10666	0.1	0	Scarce
370	Sage	10633	0.1	0	Scarce
371	Modula-2	10551	0.1	0	Scarce
372	API Blueprint	10522	0.1	0	Scarce
373	Easybuild	10439	0.1	0	Scarce
374	Texinfo	10330	0.1	0	Scarce
375	Parrot Internal Representation	10266	0.1	0	Scarce
376	Ceylon	10193	0.1	0	Scarce
377	Red	10159	0.1	0	Scarce
378	Rascal	9949	0.1	0	Scarce
379	Brainfuck	9854	0.1	0	Scarce
380	NSIS	9820	0.1	0	Scarce
381	Ninja	9600	0.1	0	Scarce
382	Monkey C	9082	0.1	0	Scarce
383	Open Policy Agent	9028	0.1	0	Scarce
384	JSON5	8962	0.1	0	Scarce
385	Modula-3	8861	0.1	0	Scarce
386	Prisma	8835	0.1	0	Scarce
387	Literate CoffeeScript	8619	0.1	0	Scarce
388	BlitzMax	8557	0.1	0	Scarce
389	Debian Package Control File	8552	0.1	0	Scarce
390	ImageJ Macro	8551	0.1	0	Scarce
391	Readline Config	8308	0.1	0	Scarce
392	Bicep	8274	0.1	0	Scarce
393	Boo	8238	0.1	0	Scarce

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
394	GAMS	8195	0.1	0	Scarce
395	Csound Score	8182	0.1	0	Scarce
396	JFlex	8105	0.1	0	Scarce
397	Genero	8084	0.1	0	Scarce
398	Microsoft Developer Studio Project	8080	0.1	0	Scarce
399	Genero Forms	8053	0.1	0	Scarce
400	J	8032	0.1	0	Scarce
401	Xojo	7973	0.1	0	Scarce
402	CUE	7867	0.1	0	Scarce
403	GDB	7784	0.1	0	Scarce
404	Alloy	7764	0.1	0	Scarce
405	Bluespec	7624	0.1	0	Scarce
406	BlitzBasic	7612	0.1	0	Scarce
407	Metal	7483	0.1	0	Scarce
408	P4	7453	0.1	0	Scarce
409	TLA	7303	0.1	0	Scarce
410	Ring	7299	0.1	0	Scarce
411	STAR	7210	0.1	0	Scarce
412	Monkey	7192	0.1	0	Scarce
413	Boogie	7145	0.1	0	Scarce
414	RPC	7024	0.1	0	Scarce
415	NASL	7003	0.1	0	Scarce
416	Macaulay2	6970	0.1	0	Scarce
417	Pony	6928	0.1	0	Scarce
418	Oz	6774	0.1	0	Scarce
419	LTspice Symbol	6717	0.1	0	Scarce
420	Singularity	6669	0.1	0	Scarce
421	Browserslist	6634	0.1	0	Scarce
422	ATS	6633	0.1	0	Scarce
423	Fantom	6622	0.1	0	Scarce
424	Genshi	6608	0.1	0	Scarce
425	APL	6463	0.1	0	Scarce
426	Berry	6463	0.1	0	Scarce
427	Dafny	6224	0.1	0	Scarce
428	SSH Config	6136	0.1	0	Scarce
429	Terra	6090	0.1	0	Scarce
430	Nemerle	6064	0.1	0	Scarce
431	LSL	6056	0.1	0	Scarce
432	ASN.1	6013	0.1	0	Scarce
433	Asymptote	6010	0.1	0	Scarce
434	Witcher Script	5987	0.1	0	Scarce
435	Zephir	5890	0.1	0	Scarce
436	Uno	5804	0.1	0	Scarce
437	2-Dimensional Array	5766	0.1	0	Scarce
438	RenderScript	5751	0.1	0	Scarce
439	E	5749	0.1	0	Scarce
440	ECL	5748	0.1	0	Scarce
441	EmberScript	5724	0.1	0	Scarce
442	Closure Templates	5713	0.1	0	Scarce
443	Brightscript	5681	0.1	0	Scarce
444	REXX	5660	0.1	0	Scarce
445	Csound	5653	0.1	0	Scarce
446	SugarSS	5423	0.1	0	Scarce
447	Zeek	5378	0.1	0	Scarce
448	Inform 7	5361	0.1	0	Scarce
449	Type Language	5310	0.1	0	Scarce
450	Futhark	4925	0.0	0	Scarce
451	XS	4917	0.0	0	Scarce
452	CAP CDS	4699	0.0	0	Scarce
453	DataWeave	4682	0.0	0	Scarce
454	MiniYAML	4669	0.0	0	Scarce
455	OpenRC runscript	4630	0.0	0	Scarce
456	Q#	4615	0.0	0	Scarce
457	Propeller Spin	4575	0.0	0	Scarce
458	Spline Font Database	4569	0.0	0	Scarce
459	Soong	4457	0.0	0	Scarce

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
460	Dylan	4397	0.0	0	Scarce
461	XC	4396	0.0	0	Scarce
462	CartoCSS	4354	0.0	0	Scarce
463	Bikeshed	4353	0.0	0	Scarce
464	Faust	4304	0.0	0	Scarce
465	Cadence	4295	0.0	0	Scarce
466	jq	4274	0.0	0	Scarce
467	ObjDump	4223	0.0	0	Scarce
468	Move	4205	0.0	0	Scarce
469	Nearley	4202	0.0	0	Scarce
470	X10	4200	0.0	0	Scarce
471	Grammatical Framework	4193	0.0	0	Scarce
472	Altium Designer	4167	0.0	0	Scarce
473	ooc	4038	0.0	0	Scarce
474	IGOR Pro	3934	0.0	0	Scarce
475	Wren	3900	0.0	0	Scarce
476	SRecode Template	3822	0.0	0	Scarce
477	eC	3810	0.0	0	Scarce
478	Fennel	3798	0.0	0	Scarce
479	Motoko	3778	0.0	0	Scarce
480	NetLogo	3775	0.0	0	Scarce
481	EQ	3759	0.0	0	Scarce
482	Harbour	3626	0.0	0	Scarce
483	VCL	3615	0.0	0	Scarce
484	Whiley	3580	0.0	0	Scarce
485	PEG.js	3519	0.0	0	Scarce
486	STON	3419	0.0	0	Scarce
487	Hy	3344	0.0	0	Scarce
488	Beef	3288	0.0	0	Scarce
489	StringTemplate	3287	0.0	0	Scarce
490	LOLCODE	3278	0.0	0	Scarce
491	hoon	3250	0.0	0	Scarce
492	Ragel	3232	0.0	0	Scarce
493	IRC log	3184	0.0	0	Scarce
494	XProc	3182	0.0	0	Scarce
495	Pic	3154	0.0	0	Scarce
496	Odin	3070	0.0	0	Scarce
497	mupad	3003	0.0	0	Scarce
498	Jolie	2941	0.0	0	Scarce
499	ABAP CDS	2840	0.0	0	Scarce
500	Frege	2828	0.0	0	Scarce
501	REALbasic	2781	0.0	0	Scarce
502	Nu	2781	0.0	0	Scarce
503	FLUX	2775	0.0	0	Scarce
504	Pike	2742	0.0	0	Scarce
505	CIL	2710	0.0	0	Scarce
506	KakouneScript	2675	0.0	0	Scarce
507	nanorc	2673	0.0	0	Scarce
508	GEDCOM	2669	0.0	0	Scarce
509	Literate Agda	2654	0.0	0	Scarce
510	Cloud Firestore Security Rules	2600	0.0	0	Scarce
511	Kusto	2550	0.0	0	Scarce
512	C2hs Haskell	2541	0.0	0	Scarce
513	Slice	2468	0.0	0	Scarce
514	Filebench WML	2465	0.0	0	Scarce
515	Component Pascal	2465	0.0	0	Scarce
516	Chapel	2457	0.0	0	Scarce
517	KiCad Legacy Layout	2454	0.0	0	Scarce
518	SmPL	2452	0.0	0	Scarce
519	MTML	2422	0.0	0	Scarce
520	Curry	2402	0.0	0	Scarce
521	Logtalk	2358	0.0	0	Scarce
522	Qt Script	2354	0.0	0	Scarce
523	Jison	2348	0.0	0	Scarce
524	RouterOS Script	2333	0.0	0	Scarce
525	HolyC	2272	0.0	0	Scarce

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
526	EBNF	2247	0.0	0	Scarce
527	Apollo Guidance Computer	2222	0.0	0	Scarce
528	Tea	2207	0.0	0	Scarce
529	SQLPL	2204	0.0	0	Scarce
530	WebIDL	2186	0.0	0	Scarce
531	LFE	2144	0.0	0	Scarce
532	HTML+ECR	2134	0.0	0	Scarce
533	KRL	2128	0.0	0	Scarce
534	BrighterScript	2114	0.0	0	Scarce
535	Edje Data Collection	2101	0.0	0	Scarce
536	Creole	2079	0.0	0	Scarce
537	GAML	2031	0.0	0	Scarce
538	Cap'n Proto	2027	0.0	0	Scarce
539	Zimpl	2015	0.0	0	Scarce
540	NetLinx	2013	0.0	0	Scarce
541	Avro IDL	1955	0.0	0	Scarce
542	Kaitai Struct	1896	0.0	0	Scarce
543	Janet	1833	0.0	0	Scarce
544	UrWeb	1818	0.0	0	Scarce
545	Talon	1772	0.0	0	Scarce
546	PicoLisp	1758	0.0	0	Scarce
547	MUF	1745	0.0	0	Scarce
548	HyPhy	1729	0.0	0	Scarce
549	Gemini	1728	0.0	0	Scarce
550	ZIL	1713	0.0	0	Scarce
551	DenizenScript	1704	0.0	0	Scarce
552	Gleam	1678	0.0	0	Scarce
553	Valve Data Format	1672	0.0	0	Scarce
554	ECLiPSe	1652	0.0	0	Scarce
555	Pod 6	1623	0.0	0	Scarce
556	Regular Expression	1601	0.0	0	Scarce
557	JSONiq	1589	0.0	0	Scarce
558	Muse	1582	0.0	0	Scarce
559	Antlers	1568	0.0	0	Scarce
560	Riot	1537	0.0	0	Scarce
561	Mint	1519	0.0	0	Scarce
562	XPages	1496	0.0	0	Scarce
563	Cycript	1482	0.0	0	Scarce
564	Pep8	1461	0.0	0	Scarce
565	Git Revision List	1426	0.0	0	Scarce
566	Elvish	1425	0.0	0	Scarce
567	Xonsh	1386	0.0	0	Scarce
568	Click	1358	0.0	0	Scarce
569	Nasal	1280	0.0	0	Scarce
570	Grace	1277	0.0	0	Scarce
571	Rouge	1252	0.0	0	Scarce
572	NL	1233	0.0	0	Scarce
573	Module Management System	1229	0.0	0	Scarce
574	Gentoo Eclass	1121	0.0	0	Scarce
575	Vyper	1114	0.0	0	Scarce
576	Win32 Message File	1109	0.0	0	Scarce
577	Cirru	1075	0.0	0	Scarce
578	Unified Parallel C	1063	0.0	0	Scarce
579	Parrot Assembly	1062	0.0	0	Scarce
580	Lark	1040	0.0	0	Scarce
581	Golo	1033	0.0	0	Scarce
582	Self	969	0.0	0	Scarce
583	dircolors	936	0.0	0	Scarce
584	X Font Directory Index	933	0.0	0	Scarce
585	Moocode	927	0.0	0	Scarce
586	Redcode	862	0.0	0	Scarce
587	ZAP	850	0.0	0	Scarce
588	Befunge	846	0.0	0	Scarce
589	Quake	838	0.0	0	Scarce
590	Mirah	813	0.0	0	Scarce
591	Ox	789	0.0	0	Scarce

Continued on next page

Rank	Language	File Count	Tokens (Billions)	Resource_Tier	Tier_Name
592	Nit	784	0.0	0	Scarce
593	Fancy	761	0.0	0	Scarce
594	TextMate Properties	760	0.0	0	Scarce
595	PogoScript	757	0.0	0	Scarce
596	HOCON	738	0.0	0	Scarce
597	Yul	729	0.0	0	Scarce
598	LigoLANG	701	0.0	0	Scarce
599	LoomScript	675	0.0	0	Scarce
600	ABNF	637	0.0	0	Scarce
601	Opa	635	0.0	0	Scarce
602	Ioke	618	0.0	0	Scarce
603	Augeas	601	0.0	0	Scarce
604	Mask	580	0.0	0	Scarce
605	Scaml	570	0.0	0	Scarce
606	Formatted	561	0.0	0	Scarce
607	CameLIGO	559	0.0	0	Scarce
608	Dogescript	542	0.0	0	Scarce
609	GCC Machine Description	534	0.0	0	Scarce
610	Shen	532	0.0	0	Scarce
611	Isabelle ROOT	531	0.0	0	Scarce
612	Cairo	528	0.0	0	Scarce
613	RUNOFF	527	0.0	0	Scarce
614	XCompose	506	0.0	0	Scarce
615	cURL Config	490	0.0	0	Scarce
616	Clarity	458	0.0	0	Scarce
617	FIGlet Font	453	0.0	0	Scarce
618	TXL	452	0.0	0	Scarce
619	wisp	370	0.0	0	Scarce
620	Wget Config	366	0.0	0	Scarce
621	Glyph	365	0.0	0	Scarce
622	Filterscript	358	0.0	0	Scarce
623	Charity	353	0.0	0	Scarce
624	Genie	313	0.0	0	Scarce
625	Sieve	306	0.0	0	Scarce
626	Bison	256	0.0	0	Scarce
627	ShellCheck Config	245	0.0	0	Scarce
628	ReasonLIGO	233	0.0	0	Scarce
629	Earthly	200	0.0	0	Scarce
630	Jison Lex	188	0.0	0	Scarce
631	Opal	164	0.0	0	Scarce
632	TI Program	163	0.0	0	Scarce
633	Darcs Patch	109	0.0	0	Scarce
634	ShellSession	104	0.0	0	Scarce
635	Oxygene	76	0.0	0	Scarce
636	Parrot	64	0.0	0	Scarce
637	Myghty	59	0.0	0	Scarce
638	Object Data Instance Notation	40	0.0	0	Scarce
639	Ecere Projects	34	0.0	0	Scarce
640	NumPy	26	0.0	0	Scarce
641	Omgrofl	17	0.0	0	Scarce
642	Python traceback	9	0.0	0	Scarce
643	MiniD	5	0.0	0	Scarce
644	NetLinx+ERB	5	0.0	0	Scarce
645	Record Jar	4	0.0	0	Scarce
646	C-ObjDump	3	0.0	0	Scarce

What Persists across Web Crawls? Genre and News Topic Variation in South Slavic Web Corpora

Nikola Ljubešić[1,2,3], Vuk Dinić[1,4], Peter Rupnik[1,4]
[1]Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia,
[2]Institute of Contemporary History, Privoz 11, Ljubljana, Slovenia,
[3]Faculty of Computer and Information Science, Večna pot 113, Ljubljana, Slovenia,
[4]Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia,
Correspondence: nikola.ljubesic@ijs.si

Abstract

This work investigates content persistence across two independently constructed versions of the CLASSLA-web corpora for seven South Slavic languages, crawled approximately two years apart, with available document-level genre and topic metadata. Persistence is defined as the presence of a near-duplicate document in both versions of the corpora, which occurs on average in 19% of cases. Genre and news-topic persistence rankings are moderately consistent across languages. Web domain availability after two years is moderately associated with genre persistence ($r = .49$, $R^2 = .24$), but not with topic persistence within news ($r = -.04$). Topic-level differences therefore appear to arise mainly within surviving news websites, where reference-oriented content tends to persist more than event-bound content.

1 Introduction

Large-scale web corpora have become a central resource for natural language processing (NLP), particularly for languages for which manually curated data remain limited (Baroni et al., 2009; Burchell et al., 2025). This is especially relevant for South Slavic languages, where web-derived corpora support both linguistic research and the development of language technologies, including language models, machine translation systems, and downstream NLP applications (Popović, 2020; Ljubešić and Lauc, 2021).

A characteristic feature of web data is the substantial amount of duplicate and near-duplicate content. When the same language-specific web space is crawled repeatedly, some documents are retrieved again, while others disappear, change, or are no longer rediscovered by the crawling process. Successive corpora therefore represent partially overlapping snapshots whose differences reflect both changes on the web and the behaviour of the corpus-construction pipeline. Although duplicate content is usually treated as noise and removed during preprocessing, overlap between crawls can also serve as a signal of content persistence, which can inform crawling frequency, crawling foci, etc.

In this paper, we investigate such persistence across two independently constructed versions of the CLASSLA-web corpora (Ljubešić and Kuzman, 2024; Kuzman Pungeršek et al., 2026) for seven South Slavic languages, crawled approximately two years apart using comparable crawling and processing procedures. Our aim is not to estimate the persistence of the web as a whole. Rather, taking Version 2 documents as the reference population, we examine which types were also represented by a near-duplicate in Version 1 under a largely unchanged corpus-construction setting.

More specifically, we analyse how document persistence varies across automatically assigned genres (Kuzman et al., 2023) and topics (Kuzman and Ljubešić, 2025), whether these patterns are consistent across languages, and to what extent they can be associated with the continued availability of the source websites. In doing so, we treat cross-crawl duplication not merely as a preprocessing problem, but as evidence about the differential persistence and turnover of web content.

2 Related Work

South Slavic web corpora. Web corpora have long provided an important source of data for South Slavic languages. Early resources such as *slWaC*, *hrWaC*, *srWaC*, and *bsWaC* followed the Web-as-Corpus methodology (Baroni et al., 2009; Ljubešić and Erjavec, 2011; Ljubešić and Klubička, 2014), while large multilingual collections such as *CC100*, *mC4*, and *HPLT* later expanded the amount of available web data (Conneau et al., 2020; Raffel et al., 2020; Burchell et al., 2025). The *CLASSLA-web* corpora provide a more curated and directly

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 53–59
October 29, 2026 ©2026 Association for Computational Linguistics

comparable collection for South Slavic languages, constructed using a shared crawling and processing pipeline (Ljubešić and Kuzman, 2024). Their two independently collected versions therefore offer a suitable setting for studying overlap between crawls.

Cross-crawl deduplication during the preparation of South Slavic language-model training data has already suggested substantial variation in such overlap. Deduplication among the earlier WaC corpora removed only single-digit percentages of the data used for pre-training the *BERTić* encoder model, while approximately 15% of additional Common Crawl material overlapped with those corpora (Ljubešić and Lauc, 2021). When newer corpora were used to construct the *XLM-R-BERTić* training dataset and the related continuously pretrained multilingual encoder model, only 45% of the earlier *BERTić* data remained after deduplication (Ljubešić et al., 2024). Most recently, and most relevant for this work, approximately 20% of documents in the second version of *CLASSLA-web* were matched to documents in the first version (Kuzman Pungeršek et al., 2026). These observations show that independently collected web corpora overlap only partially, but do not explain which types of content are most likely to recur.

Web persistence and change. Research on web persistence predates current large-scale NLP corpora. Early longitudinal studies estimated an average web-page lifetime of approximately two years and documented continual change in both pages and websites (Koehler, 1999, 2002). Later work showed that change is highly uneven: frequently visited pages change more often than randomly sampled pages, and content survival follows a characteristic J-shaped decay pattern (Adar et al., 2009). Studies of reference rot and long-term web accessibility similarly demonstrate that substantial portions of previously available web content eventually become unreachable (Klein et al., 2014; Chapekis et al., 2024).

These studies generally examine the continued accessibility or modification of known URLs. Cross-crawl overlap measures a related but different phenomenon. A document is observed as persistent only if sufficiently similar content is retrieved and retained by both crawling and processing pipelines. Failure to observe it again may therefore result from page removal, content change, website unavailability, crawl coverage, or filtering decisions. In this paper, we use *persistence* in this operational, corpus-level sense and subsequently examine whether differences in cross-crawl overlap can be partly explained by the later availability of source websites.

Duplicate as signal. In contemporary NLP, duplicate web content is normally treated as undesirable redundancy. Deduplication reduces memorisation and training cost without necessarily harming model quality (Lee et al., 2022), and large web datasets may discard substantial portions of their input through exact and near-duplicate removal (Penedo et al., 2023). In contrast, we use near-duplicate matches as evidence of content recurring across independently constructed corpora. Our main contribution is to relate the content recurrence to automatically assigned genre and topic labels across seven South Slavic languages, thereby investigating whether some categories of web content are systematically more persistent than others. These insights can inform the design of future crawls, as well as support retrospective interpretation of existing web corpora, in particular by showing which types of content are more likely to be missed or underrepresented in existing crawls.

3 Data and Research Questions

Our analysis uses the two versions of the CLASSLA-web corpora (Ljubešić and Kuzman, 2024; Kuzman Pungeršek et al., 2026), which cover seven South Slavic languages. Version 1 was crawled in 2021–2022 and consists of 26 million documents and 11 billion words, and Version 2 in 2024 containing 38 million documents and 17 billion words, both versions using comparable crawling and processing pipelines. Near-duplicates were removed within each version independently.

During the construction of Version 2, documents were also matched across versions using MinHash (Shrivastava and Li, 2014) configured with an LSH similarity threshold of 0.7, 96 MinHash permutations, and four-word shingles. This particular MinHash setup was identified via manual inspection on multiple iterative runs over these particular data collections (Kuzman Pungeršek et al., 2026).

We treat a Version 2 document as *persistent* if a near-duplicate was found in Version 1. Thus, persistence is directional: it is the proportion of documents in the later corpus that are also represented in the earlier corpus. Also, this is an operational measure of recurrence across independent crawls

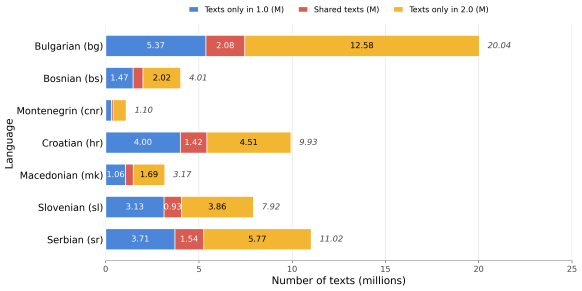


Figure 1: Numbers of documents in CLASSLA-web Versions 1 and 2 and their cross-version overlap by language, adapted from Kuzman Pungeršek et al. (2026).

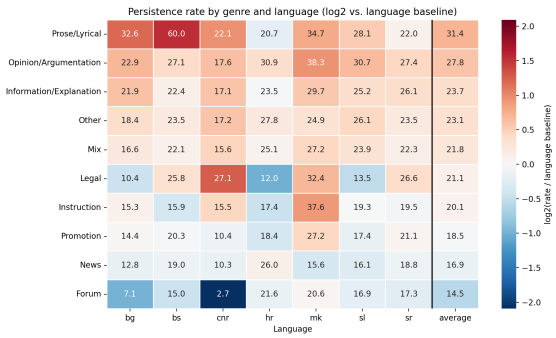


Figure 2: Persistence rates by genre and language. A Version 2 document is considered persistent if a near-duplicate is found in Version 1. *Mix* is a low-confidence fallback class.

rather than direct evidence that a particular URL remained unchanged. Across languages, 12–24% of Version 2 documents were matched to Version 1, with an average persistence rate of 19%. Figure 1 summarizes corpus sizes and overlap by language.

We examine whether persistence varies systematically across the genres and topics automatically assigned to Version 2 documents, and whether the same patterns recur across languages. Topic-level analysis is conducted within the most prominent news genre, for which the topic classifier was developed and evaluated. Finally, we test whether differences in persistence can be partly explained by the later availability of the corresponding websites and documents.

4 Results

4.1 Genre Persistence

To contextualize our results on genre persistence, we first estimate the overall genre distribution from equally sized samples (10 thousand documents) of the seven Version 2 corpora. News accounts for 56% of documents, followed by Promotion (10%), Forum and Information/Explanation (8% each), Opinion/Argumentation (6%), Instruction (4%), Mix (3%), Other (2%), and Prose/Lyrical and Legal (1% each). It is important to distinguish between *Other* and *Mix* categories. *Other* is an explicit category used during manual annotation for documents that do not fit any category in the annotation schema, whereas *Mix* is assigned automatically at inference time when no genre reaches the confidence threshold of 0.8 (Kuzman Pungeršek et al., 2026).

Figure 2 presents persistence rates by genre and language. Cell values give the raw persistence percentages, while colour indicates the $\log_2$ ratio between the genre-specific rate and the overall persistence baseline for that language: red denotes above-baseline persistence, blue below-baseline persistence, and near-white values indicate little difference from the baseline. The results show substantial and moderately consistent differences among genres. Prose/Lyrical content is generally the most persistent, whereas Forum content is among the least persistent. To assess whether the relative ordering of genres is shared across languages, we calculated Spearman rank correlations for all 21 language pairs. The median correlation was $\rho = .53$, indicating moderate cross-language consistency. Pairwise values ranged from $\rho = -.24$ for Croatian and Macedonian to $\rho = .94$ for Bulgarian and Slovenian, with Croatian showing the least consistent pattern overall.

4.2 Topic Persistence

We again start the discussion on persistence of topics with an insight into the most frequent topics in the language-balanced sample of Version 2, most frequent topics being Politics, Arts, Culture, Entertainment and Media, and Economy, Business and Finance (12% each), followed by Sport (11%), Human Interest (8%), and Mix and Lifestyle and Leisure (7% each). The remaining topics individually account for 5% or less of the data. For topic classification, *Mix* is assigned automatically when no single topic reaches the confidence threshold of 0.6; it therefore denotes documents without a confidently identifiable dominant topic rather than a manually defined substantive category (Kuzman Pungeršek et al., 2026).

Because the IPTC topic classifier was developed and evaluated on news data, we report topic persistence within the News genre, which also constitutes

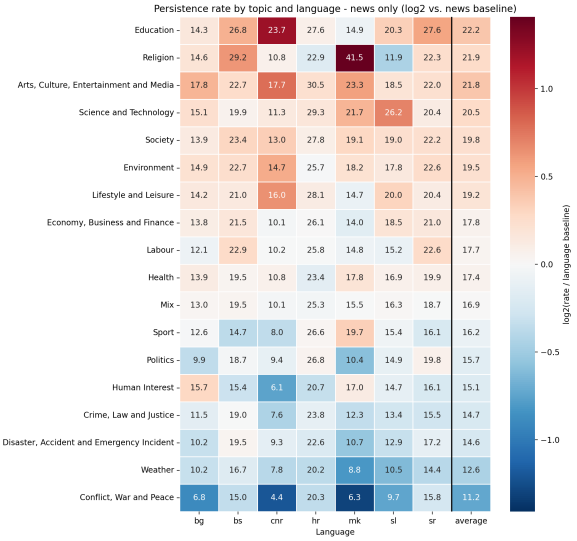


Figure 3: Persistence rates by news topic and language. *Mix* is a low-confidence fallback class.

more than half of the overall content.

The ordering of news topics is moderately to strongly consistent across languages. The median pairwise Spearman correlation is $\rho = .55$, and all 21 language-pair correlations are positive, ranging from $\rho = .43$ to $\rho = .92$. Topic rankings are therefore somewhat more uniformly shared across languages than genre rankings.

The most persistent topics tend to contain reusable or reference-oriented content, including Education, Religion, Arts and Culture, and Science and Technology. The least persistent topics are more event-bound, particularly Conflict, Weather, Disasters, and Crime.

4.3 Persistence and Web Domain Availability

A document may fail to recur across crawls because its source website has disappeared, but also because the page has changed, moved, been removed, or was not rediscovered by the crawler. To assess the contribution of website survival, we sampled 100 Version 2 documents for each genre and each news topic in every language and tested the availability of their source web domains in July 2026, approximately two years after the Version 2 crawl, obtaining thereby genre-specific and topic-specific website availability rates, structurally comparable to the persistence rates presented previously. Average website availability remained between roughly 70% and 90%, substantially above the 12–24% cross-crawl persistence rates.

To estimate the relationship between persistence and web domain availability, we calculated Pearson correlations between persistence rates and availability rates. We additionally fitted a simple linear regression of persistence on domain availability to obtain R^2. All language–category combinations received equal weight. The complete availability rates are reported in Appendix A.

Across the 70 genre–language combinations, web domain availability was moderately associated with persistence ($r = .49$, $p < .001$, $R^2 = .24$). Genres occurring on more durable websites therefore also tended to recur more often across crawls. In contrast, no association was observed across news-topic–language combinations ($r = -.04$, $R^2 < .01$). Website survival may thus explain part of the variation among genres, whereas topic-level differences within news appear to arise mainly from processes within surviving websites. Our working hypothesis still remains that reference-oriented content (news on education, arts, religion, science) has a much slower turnover than event-bound content (news on wars, crime, disasters and weather).

5 Conclusion

We examined content persistence across two independently constructed versions of the CLASSLA-web corpora for seven South Slavic languages, collected approximately two years apart. On average, 19% of Version 2 documents had a near-duplicate in Version 1, indicating substantial turnover between otherwise comparable crawls.

Persistence varied systematically across both genre and topic. Genre rankings were moderately consistent across languages, with Prose/Lyrical content generally among the most persistent and Forum content among the least. Topic-level patterns within news were more uniform: reusable and reference-oriented topics tended to persist more, whereas event-bound topics such as Conflict, Weather, Disasters, and Crime were less persistent.

Website survival explained only part of these differences. Web domain availability was moderately associated with genre persistence, but unrelated to persistence across news topics. Much of the observed news turnover therefore appears to arise within surviving websites.

Future work should exploit the released data[1] primarily to disentangle the mechanisms underlying cross-crawl persistence.

[1]The derived TSV-formatted dataset is available at http://hdl.handle.net/11356/1972.

Limitations

Our measure of persistence captures recurrence across two independently constructed corpora rather than direct survival of individual web pages. Documents may fail to match because they were created only after the first crawl, were removed or substantially changed, were not rediscovered by the crawler, were filtered during processing, or fell below the near-duplicate threshold. The latter may also affect genres differently depending on document length and structure. Accordingly, the observed rates should be interpreted as cross-crawl recurrence rather than direct estimates of content lifetime or turnover.

Genre and topic labels were assigned automatically and therefore contain classification errors. Topic analysis is restricted to news because the classifier was primarily evaluated in that genre.

The availability analysis is based on samples of 100 documents per category and language and measures domain reachability only at a single follow-up point in July 2026, rather than during the interval between the two crawls. It therefore provides only an indirect proxy for website survival, particularly for news topics, which largely share domains and show limited variation in availability. Finally, the study covers two crawl waves, seven closely related languages, and one corpus-construction and near-duplicate matching setup, limiting generalizability to other web collections and time intervals.

Acknowledgements

This paper was produced as part of the project "Large Language Models for Digital Humanities" (Grant GC-0002), the research programme "Language Resources and Technologies for Slovene" (Grant P6-0411), and the Research Infrastructure DARIAH-SI (I0-E007), all funded by the ARIS Slovenian Research and Innovation Agency, as well as a part of the Datavysts MSCA network, funded by Horizon Europe (grant number 101169037) and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding Guarantee (grant number EP/Z536581/1).

This research was also supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. This project is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

Eytan Adar, Jaime Teevan, Susan T. Dumais, and Jonathan L. Elsas. 2009. The web changes everything: Understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 282–291, Barcelona Spain. ACM.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Laurie Burchell, Ona De Gibert Bonet, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 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 (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Athena Chapekis, Samuel Bestvater, Emma Remy, and Gonzalo Rivero. 2024. When online content disappears.

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, Online. Association for Computational Linguistics.

Martin Klein, Herbert Van De Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE*, 9(12):e115253.

Wallace Koehler. 1999. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2):162–180.

Wallace Koehler. 2002. Web page change and persistence—a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2):162–171.

Taja Kuzman and Nikola Ljubešić. 2025. LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification. *IEEE Access*, pages 1–1.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman Pungeršek, Peter Rupnik, Vít Suchomel, and Nikola Ljubešić. 2026. The growing gains and pains of iterative web corpora crawling: Insights from South Slavic CLASSLA-web 2.0 corpora. *arXiv preprint arXiv:2601.11170*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for croatian and slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. Springer.

Nikola Ljubešić and Filip Klubička. 2014. {bs, hr, sr} wac-web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th web as corpus workshop (WaC-9)*, pages 29–35.

Nikola Ljubešić and Taja Kuzman. 2024. CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3271–3282, Torino, Italia. ELRA and ICCL.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Nikola Ljubešić, Vít Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024. Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 189–203, Torino, Italia. ELRA and ICCL.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36.

Maja Popović. 2020. Neural machine translation for translating into croatian and serbian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 102–113. International Committee on Computational Linguistics.

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.

Anshumali Shrivastava and Ping Li. 2014. In defense of minhash over simhash. In *Artificial intelligence and statistics*, pages 886–894. PMLR.

A Domain availability by language and genre/topic.

In Table 1 we present full availability rates (percentage of websites that were still available two years after the Version 2 crawl, as described in Section 4.3) for individual genres and languages. For documents with the genre label News, we further present their availability rates by their topic and language in Table 2.

Genre	bg	bs	cnr	hr	mk	sl	sr	Average
Prose/Lyrical	0.93	0.97	0.93	0.92	0.96	0.86	0.93	0.93
Other	0.89	0.87	0.86	0.90	0.93	0.87	0.89	0.89
Instruction	0.89	0.91	0.88	0.76	0.97	0.86	0.84	0.87
Information/Explanation	0.87	0.92	0.82	0.80	0.92	0.88	0.89	0.87
Mix	0.81	0.85	0.91	0.83	0.96	0.89	0.85	0.87
Opinion/Argumentation	0.85	0.75	0.91	0.80	0.98	0.94	0.82	0.86
News	0.82	0.87	0.85	0.85	0.81	0.96	0.87	0.86
Legal	0.95	0.85	0.80	0.42	0.96	0.83	0.91	0.82
Promotion	0.92	0.89	0.68	0.72	0.93	0.65	0.86	0.81
Forum	0.73	0.83	0.21	0.91	0.99	0.92	0.87	0.78

Table 1: Web domain availability rate by genre and language.

Topic	bg	bs	cnr	hr	mk	sl	sr	Average
Weather	0.82	0.90	0.94	0.93	0.93	0.92	0.92	0.91
Disaster, Accident and Emergency Incident	0.83	0.92	0.91	0.92	0.87	0.96	0.92	0.90
Labour	0.88	0.88	0.91	0.92	0.90	0.91	0.92	0.90
Science and Technology	0.93	0.86	0.91	0.91	0.93	0.91	0.85	0.90
Conflict, War and Peace	0.84	0.87	0.91	0.89	0.93	0.94	0.85	0.89
Environment	0.77	0.88	0.92	0.88	0.87	0.97	0.92	0.89
Mix	0.80	0.85	0.94	0.91	0.87	0.95	0.87	0.88
Lifestyle and Leisure	0.81	0.90	0.92	0.88	0.86	0.94	0.88	0.88
Arts, Culture, Entertainment and Media	0.88	0.85	0.91	0.88	0.83	0.96	0.85	0.88
Crime, Law and Justice	0.84	0.84	0.91	0.88	0.87	0.95	0.90	0.88
Human Interest	0.85	0.82	0.92	0.87	0.81	0.97	0.90	0.88
Health	0.71	0.87	0.95	0.91	0.88	0.94	0.88	0.88
Religion	0.81	0.95	0.91	0.85	0.78	0.93	0.88	0.87
Politics	0.85	0.84	0.92	0.87	0.88	0.88	0.81	0.86
Education	0.83	0.89	0.94	0.87	0.82	0.73	0.91	0.86
Society	0.75	0.83	0.93	0.90	0.80	0.92	0.88	0.86
Sport	0.90	0.92	0.88	0.72	0.72	0.94	0.91	0.86
Economy, Business and Finance	0.84	0.80	0.89	0.86	0.77	0.95	0.83	0.85

Table 2: Web domain availability by news topic and language.

HPLT3-PerRef: A Refined Large-Scale Persian Web Corpus via Supervised Document Quality Filtering

Alireza Razzaghi, Tomasz Galica, Erik Henriksson, Veronika Laippala
TurkuNLP, University of Turku
{alireza.razzaghi, tomasz.galica, erik.henriksson, mavela}@utu.fi

Abstract

Large-scale web corpora are valuable resources for natural language processing (NLP), linguistic research, and language model training, particularly for digitally low-resource languages such as Persian. However, web-derived text often contains document-level noise, even after quality-filtering pipelines. We present HPLT3 Persian Refined (HPLT3-PerRef), a refined Persian web corpus containing 105.9 billion tokens and 102.37 million documents, obtained by applying a supervised quality-filtering pipeline to the Persian subset of HPLT3. We manually annotated 6,144 Persian web documents and trained a multilingual transformer-based classifier to distinguish high- and low-quality documents. The classifier removed approximately 20% of the original corpus while preserving broad domain coverage and complements HPLT3's heuristic-based Web Document Score (WDS) by identifying additional low-quality documents. We publicly release the binary annotated corpus, the trained classifier, the HPLT3-PerRef, and the complete source code on GitHub, Hugging Face (Data) and Hugging Face (Quality classifier).

1 Introduction

The web has become a valuable source of textual data for linguistic research and natural language processing (NLP), providing large-scale, diverse, and continuously updated language data that are difficult to obtain from curated resources alone. This is particularly important for digitally low-resource languages such as Persian, where high-quality text collections remain limited. However, web-derived corpora often contain document-level noise that can affect linguistic analyses and language model training, motivating continued efforts to improve corpus quality.

Several monolingual Persian (Farsi) corpora have been developed in recent years (Hosseinbeigi et al., 2025; Khansari and Shamsfard, 2021; Ayoubi, 2022; Sarmadi et al., 2025; Sabouri et al., 2024; Ayoubi, 2021; Darvishi et al., 2023; Sabeti et al., 2018; Doostmohammadi et al., 2018; AleAhmad et al., 2016, 2009; Shojaei, 2024). However, many are limited in scale, are not openly accessible, or focus on specific content types such as news or blogs. Large multilingual web corpora (Nguyen et al., 2024; Oepen et al., 2025; Sagot and Romary, 2019; Xue et al., 2021; Kudugunta et al., 2023) provide broader Persian coverage but also include noisy documents, such as boilerplate, SEO-oriented pages, machine-translated text, and archive pages. Such documents increase corpus size while reducing its value for linguistic research and language model training.

We present HPLT3-PerRef, a 105.9-billion-token Persian corpus derived from the Persian portion of HPLT3 (Oepen et al., 2025). Starting from HPLT3's heuristic-filtered corpus, we further refine the Persian subset using supervised document-level quality classification. Despite this additional filtering, HPLT3-PerRef remains among the largest publicly available Persian corpora (Appendix Table 4). Specifically, we (i) manually annotate a subset of HPLT3 documents to train a document quality classifier, (ii) apply the classifier to the full HPLT3 Persian part to remove low-quality documents and assign quality confidence scores to the retained ones, and (iii) publicly release the refined corpus, the classifier, the document-level quality scores, and the source code. Finally, we show that the classifier complements HPLT's Web Document Score (WDS) by identifying additional low-quality documents.

2 Related Work

Existing Persian language resources span a wide range of domains and purposes. Large-scale web corpora include Matina (Hosseinbeigi et al., 2025),

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 60–69
October 29, 2026 ©2026 Association for Computational Linguistics

Targoman (Ayoubi, 2022), and MirasText (Sabeti et al., 2018). Domain-specific datasets cover blogs (HmBlogs Khansari and Shamsfard 2021, irBlogs AleAhmad et al. 2016), news (PerKey Doostmohammadi et al. 2018, Hamshahri AleAhmad et al. 2009), and PDF documents (PDC Shojaei 2024). Multi-source corpora such as Hakim (Sarmadi et al., 2025) and naab (Sabouri et al., 2024) integrate web text with resources including Wikipedia, Common Crawl, and social media. Moreover, PersianQA (Ayoubi, 2021) and PQuAD (Darvishi et al., 2023) provide task-specific benchmarks for question answering. However, many existing monolingual resources are limited in scale, focus on specific domains or tasks, or face accessibility and reproducibility challenges. For example, the largest publicly released monolingual Persian corpus, Matina, contains 72.9B tokens (Appendix 4), making it smaller than HPLT3-PerRef. In addition, access to the corpus requires approval through Hugging Face, which limits ease of access and reproducibility.

Furthermore, open-source large multilingual web corpora provide broader coverage of Persian. Representative examples are CultureX (Nguyen et al., 2024), HPLT3 (Oepen et al., 2025), OSCAR (Sagot and Romary, 2019), mC4 (Xue et al., 2021), and Madlad-400 (Kudugunta et al., 2023), which include Persian alongside many other languages. Among these, HPLT3 provides one of the largest publicly available collections of Persian web text and has a document-level filtering pipeline.

Despite these existing filtering efforts, web corpora often retain document-level noise. HPLT3 employs the WDS, a heuristic-based scoring method that estimates document quality from structural and statistical properties of web pages. Other approaches, including Propella (Idahl et al., 2026) and WebOrganizer (Wettig et al., 2025), use learned document quality models for large-scale web corpus filtering; however, they have not been developed or evaluated for Persian. These observations motivate our use of HPLT3 as the foundation for HPLT3-PerRef: it contains one of the largest publicly available collections of Persian web text, yet its heuristic WDS filtering still retains some document-level noise. HPLT3-PerRef therefore further refines the Persian subset using a supervised document-level quality classifier.

3 Corpus Construction

3.1 Source Corpus

Our corpus is constructed from the Persian portion of HPLT3. Each document corresponds to the cleaned textual content extracted from a single web page. The Persian subset contains approximately 127.96 million documents and 157.76 billion tokens, collected from the Internet Archive (2012–2020) and 57 Common Crawl snapshots (2014–2025). HPLT3 applies several large-scale processing steps, including language identification, quality filtering using the WDS, and global near-duplicate removal with MinHash, resulting in broad domain coverage and reduced duplication.

Although these preprocessing steps provide a strong foundation, additional Persian-specific quality refinement can further improve the suitability of the corpus for linguistic research and language model training. Through repeated manual inspections of randomly sampled documents from the Persian portion of HPLT3, we observed that some low-quality content remains after the existing filtering process. This suggests that heuristic quality measures may not capture all types of document-level noise.

3.2 Annotation

Building on the preprocessing performed by HPLT3, we further refine its Persian subset using a Persian-specific document quality classifier trained on a manually annotated dataset. Documents were assigned one of two labels based on their overall quality: high-quality (class 1) or low-quality (class 0).

High-quality documents were defined as coherent and complete texts forming a unified document. Low-quality documents included incomplete texts (e.g., fewer than two complete sentences), incoherent content, documents dominated by boilerplate, special characters, numerical sequences, headline fragments, lists, download pages, image captions, keyword lists, repeated search terms, link collections, or SEO-oriented content. Examples of high- and low-quality documents are shown in Figure 1.

3.3 Quality Classifier

The binary document quality classifier was trained on an annotated dataset of 6,144 Persian web documents (3,998 high-quality and 2,146 low-quality) sampled from the HPLT3 Persian corpus, and divided into training, validation, and test sets using

Low-quality: SEO page

-فیلم ملکه -خشک کردن سبزی -دانلود بازی تخته نرد
Translation: -Queen's Movie -Drying vegetables -Download backgammon game.

Low-quality: Boilerplate

صفحه اصلی | تماس با ما | قوانین
Translation: Home | Contact Us | Terms

High-quality

تیم استقلال ساری بدون حامی مالی و به عنوان تنها نماینده از مازندران وارد رقابتهای لیگ برتر فوتسال بانوان شده و در 6 مسابقه ابتدایی خود پنج بار متحمل شکست شده است.
Translation: Esteghlal Sari team, without any financial support and as the only representative from Mazandaran, entered the competition of the women's futsal league and lost five out of their first six matches.

Figure 1: Examples of manually annotated high- and low-quality documents.

an 80/10/10 split.

We evaluated two multilingual transformer encoders, XLM-RoBERTa-large (Conneau et al., 2020) and BGE-M3 (Chen et al., 2024), which are suitable for Persian web data that contains diverse linguistic patterns, including non-standard writing, code-switching, multilingual fragments, and other web-specific noise. XLM-RoBERTa was selected as a strong multilingual baseline, while BGE-M3 was chosen for its longer input sequence length (1,024 vs. 512 tokens), enabling more comprehensive document-level quality estimation. Both models were fine-tuned as binary sequence classifiers, with hyperparameters optimized using Bayesian optimization with Optuna on the training and validation sets. The search space, optimization procedure, and selected hyperparameter configuration are described in Appendix A.2. The best configuration was retrained from scratch and evaluated on the held-out test set.

Metric	BGE-M3	XLM-R
Accuracy	96.91%	97.56%
Macro F1	96.58%	97.31%
Weighted F1	96.90%	97.56%
Total errors	19/615	15/615

Table 1: Performance of quality classification models on the test set.

Although XLM-RoBERTa achieved slightly higher classification performance (Table 1), we selected BGE-M3 for large-scale corpus filtering because our goal was not only binary classification but also reliable document-level quality scoring across a continuous range. Its 1,024-token input capacity allows it to evaluate a larger portion of each web page, producing more representative quality estimates while maintaining comparable performance.

The classifier outputs a continuous score between 0 and 1 representing the predicted probability of a document belonging to the low-quality class, with higher scores indicating lower quality. Unlike a fixed binary decision, this probability-based output allows the filtering threshold to be adjusted according to the desired trade-off between corpus size and quality. We selected the final threshold on the validation set by evaluating precision and recall for the low-quality class across different thresholds, aiming to maximize low-quality document removal while retaining high-quality documents.

The BGE-M3 classifier contains approximately 567 million parameters, while XLM-RoBERTa-large contains approximately 559 million parameters. Model training was performed on a single NVIDIA A100 GPU with 256 GB of host memory and 8 CPU cores, and took approximately 3.5 hours for BGE-M3-RetroMAE and 10 minutes for XLM-RoBERTa-large. Applying the classifier to the HPLT3 data was performed on a single NVIDIA A100 GPU with 256 GB of host memory and 16 CPU cores. The HPLT3 data were processed using streaming inference across multiple jobs, with a cumulative elapsed job time of approximately 53 GPU hours.

3.4 Intrinsic Corpus-Quality Check

To provide an intrinsic validation beyond agreement with the human annotations, we conducted a controlled language-model pretraining experiment. We trained identical, randomly initialized GPT-2-style decoder-only Transformers on three corpora: HPLT3, HPLT3-PerRef, and a random-removal control created by randomly removing approximately 20% of documents from HPLT3.

The models have 6 Transformer layers, 12 attention heads, a hidden size of 768, and a context length of 2,048 tokens. We used the Qwen3.6-35B-A3B (Qwen Team, 2026) tokenizer for all three conditions, but we did not use the pre-

trained Qwen3.6-35B-A3B model or its pretrained weights. Each model was trained from a random initialization using the same initialization seed, architecture, optimization settings, and 100M-token training budget. This ensures that differences in validation loss are attributable to the training corpora rather than to differences in model configuration or training conditions.

After training, all models were evaluated on the same held-out Persian test set using next-token prediction loss and perplexity. The results are shown in Table 2. HPLT3-PerRef achieved lower held-out cross-entropy and perplexity than both the original HPLT3 corpus and the random-removal control. This result suggests that the documents removed by the supervised filtering procedure are not simply a random subset of HPLT3. Under a fixed token budget, the resulting HPLT3-PerRef corpus produces a more predictable training distribution according to this intrinsic language-modeling criterion.

Training corpus	Cross-entropy	Perplexity
HPLT3	13.14	511K
Random removal	13.22	553K
HPLT3-PerRef	12.93	414K

Table 2: Held-out language-modeling cross-entropy and perplexity on Persian test data.

3.5 HPLT3-PerRef Creation and WDS

Applying the trained BGE-M3 classifier to the Persian portion of HPLT3 produced HPLT3-PerRef, containing 102.37 million documents and approximately 105.9 billion tokens. Overall, 20% of the original documents were removed.

Because WDS relies solely on surface-level features, it may assign high scores to documents that our annotation guidelines classify as low quality. It may also assign lower scores to documents with unusual formatting, even when they contain useful linguistic content.

To investigate the relationship between WDS and our Persian-specific quality classifier, we compare the distribution of WDS scores for the entire Persian subset of HPLT3 with those of documents classified as low quality. This analysis assesses whether the proposed classifier identifies additional low-quality documents beyond those captured by WDS.

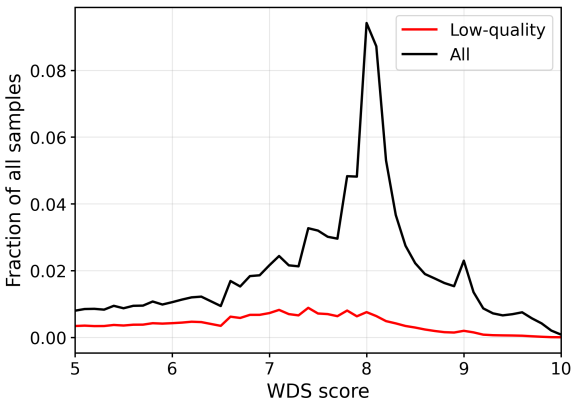


Figure 2: Distribution of WDS scores for the full HPLT3-PerRef and documents classified as low quality.

3.6 WDS analysis

Figure 2 shows the marginal distribution of WDS scores for the entire corpus and low-quality documents. Since the overall corpus is concentrated around a WDS score of approximately 8, most low-quality documents also appear in this region. However, this distribution shows the frequency of WDS scores among low-quality documents, not the probability that a document with a given WDS score is low-quality. Equivalent distributions for the remaining quality scores are in Appendix Figure 4.

Figure 3 shows the proportion of documents classified as low-quality within each WDS score bin. After normalising by the number of documents in each bin, a clear inverse relationship emerges: the proportion of low-quality documents decreases from approximately 42% at WDS = 5 to below 5% for WDS scores above 9.5. We fit a weighted four-parameter logistic (4PL) model to the observed prevalence, which provides a good fit to the empirical distribution (R2=0.967)[1]. The fitted curve captures the upper and lower plateaus and the transition region around WDS scores of 7.4–7.9, showing that the likelihood of a document being classified as low-quality decreases smoothly as WDS increases. However, low-quality documents remain present across the entire WDS score range, including the highest-scoring bins. WDS is therefore a useful quality indicator but cannot fully identify all low-quality documents, motivating the use of the proposed classifier as a complementary filtering approach. The corresponding prevalence plots for all quality metrics are in Figure 5.

[1]The weighted 4PL model outperformed quadratic ($R^2 = 0.809$) and cubic ($R^2 = 0.904$) alternatives.

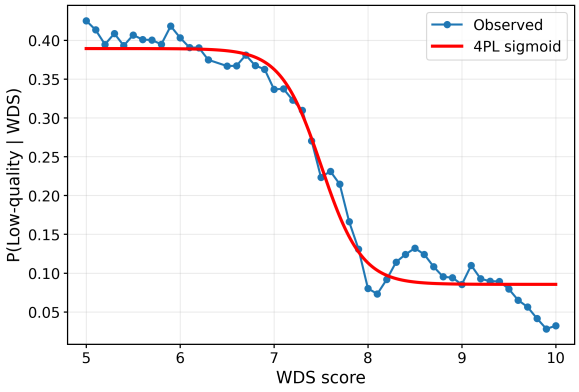


Figure 3: Proportion of documents classified as low quality across WDS score bins.

Pearson correlation analysis over the entire corpus was used to measure the linear relationship between the continuous WDS scores and the classifier's predicted low-quality probabilities. It further indicates that the proposed classifier captures information complementary to the existing HPLT document quality scores. The overall WDS score shows a moderate negative correlation with the probability of being classified as low-quality ($r=-0.279$). Among individual WDS components, correlations remain generally weak, suggesting that the classifier captures quality-related patterns that are not fully represented by the heuristic scores.

4 Conclusion

We presented HPLT3-PerRef, a refined Persian web corpus constructed from the Persian portion of HPLT3 using a supervised document quality filtering pipeline. The filtering process removed approximately 20% of the original corpus while preserving broad web coverage. Our analysis demonstrated that the proposed classifier complements HPLT3's Web Document Score (WDS) by identifying additional low-quality documents that are not fully captured by heuristic quality measures.

HPLT3-PerRef contains approximately 105.9 billion tokens across more than 102 million documents, making it one of the largest publicly available refined Persian corpora for general-purpose NLP and language model pretraining. HPLT3-PerRef and its associated resources are publicly released to support reproducible research, available at GitHub, Hugging Face (Data) and Hugging Face (Quality classifier).

Future work includes extending the annotation framework to finer-grained quality categories, incorporating additional semantic quality signals, and evaluating the impact of the refined corpus on Persian language model pretraining and downstream tasks.

5 Limitations and Further Work

The quality classifier was trained on a manually annotated subset of Persian HPLT3 documents. Since the annotations were produced by a single annotator, inter-annotator agreement could not be measured, and the labels may reflect subjective quality judgments. Furthermore, the annotated dataset represents only a subset of Persian web content, and the distribution of document types and quality levels within this sample may influence the classifier's decisions. Under-represented document types or WDS score ranges may therefore be classified less reliably.

As further work, the relationship between the heuristic WDS score and the supervised classifier could be investigated more systematically, including evaluating WDS as a baseline and exploring whether its component weights can be optimized to better reflect the annotated quality judgments. In addition, moving beyond binary quality labels toward multi-dimensional annotations could provide a more nuanced representation of the diverse quality characteristics of web documents and better capture the complexity of web data.

6 Data Disclaimer

The documents contained in HPLT3-PerRef are derived from publicly available web content and may contain opinions, claims, or other statements expressed by the original authors of those documents. Their inclusion in the corpus does not imply that the authors of this work endorse, support, or agree with any such content. The corpus is provided for research purposes, and the views and statements expressed in the included documents should not be attributed to the authors of this work.

Acknowledgments

Alireza Razzaghi received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No 101177564—HAIF. Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive

Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Tomasz Galica received funding from OpenEuroLLM and LLMs4EU projects, under Digital Europe Programme grant agreements No 101195233 and 101198470.

This work was supported by the Research Council of Finland through several projects: FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union– NextGenerationEU instrument), "Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora" (project 362459), and the "Centre of Excellence Human Diversity through Contacts" (project 374223). We also wish to acknowledge CSC– IT Center for Science Ltd. for providing computational resources

References

Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2009. Hamshahri: A standard persian text collection. *Knowledge-Based Systems*, 22(5):382–387.

Abolfazl AleAhmad, MohammadSadegh Zahedi, Maseud Rahgozar, and Behzad Moshiri. 2016. irblogs: A standard collection for studying persian bloggers. *Computers in Human Behavior*, 57:195–207.

Mohammad Yasin Ayoubi, Sajjad & Davoodeh. 2021. Persianqa: a dataset for persian question answering. `https://github.com/SajjjadAyobi/PersianQA`.

Mohammad Yasin Ayoubi, Sajjad & Davoodeh. 2022. Targoman large persian corpus. `https://huggingface.co/datasets/Targoman/TLPC`.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. Pquad: A persian question answering dataset. *Computer Speech amp; Language*, 80:101486.

Ehsan Doostmohammadi, Mohammad Hadi Bokaei, and Hossein Sameti. 2018. Perkey: A persian news corpus for keyphrase extraction and generation. In *2018 9th International Symposium on Telecommunications (IST)*, page 460–465. IEEE.

Sara Bourbour Hosseinbeigi, Fatemeh Taherinezhad, Heshaam Faili, Hamed Baghbani, Fatemeh Nadi, and Mostafa Amiri. 2025. Matina: A large-scale 73B token Persian text corpus. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9143–9157, Albuquerque, New Mexico. Association for Computational Linguistics.

Maximilian Idahl, Benedikt Droste, Björn Plüster, and Jan Philipp Harries. 2026. propella-1: Multi-property document annotation for llm data curation at scale. *Preprint*, arXiv:2602.12414.

Hamzeh Motahari Khansari and Mehrnoush Shamsfard. 2021. Hmblogs: A big general persian corpus. *Preprint*, arXiv:2111.02362.

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. *Preprint*, arXiv:2309.04662.

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, Torino, Italia. ELRA and ICCL.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Georges Gabriel Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajic, Jindrich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. HPLT 3.0: Very large-scale multilingual resources for LLM and MT. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *CoRR*, abs/2511.01066.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Behnam Sabeti, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, S.H.E. Mortazavi Najafabadi,

and Amir Vaheb. 2018. MirasText: An automatically generated text corpus for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sadra Sabouri, Elnaz Rahmati, Soroush Gooran, and Hossein Sameti. 2024. naab: A ready-to-use plug-and-play corpus for farsi. *Journal of Artificial Intelligence Applications and Innovations*, 1(2).

Benoît Sagot and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. page 8.

Mehran Sarmadi, Morteza Alikhani, Erfan Zinvandi, and Zahra Pourbahman. 2025. Hakim: Farsi text embedding model. *Preprint*, arXiv:2505.08435.

Mohammad Shojaei. 2024. Persian document corpus. `https://huggingface.co/datasets/mshojaei77/persian-document-corpus`.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. *Preprint*, arXiv:2502.10341.

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. *Preprint*, arXiv:2010.11934.

A Appendix

A.1 WDS Score

The WDS quality score is composed of several heuristic subscores. `language_score` measures the proportion of alphabetic characters belonging to the target language, `urls_score` the proportion of URLs, `punctuation_score` the proportion of punctuation characters, `singular_chars_score` the proportion of dominant single characters, `numbers_score` the proportion of numeric characters, `repeated_score` the amount of repeated text segments, `n_long_segments_score` the amount of long text segments, `great_segment_score` the length of the longest text segment, and `informativeness_score` the document's compression-based informativeness estimated using Zstandard.

Figures 4 and 5 extend the analyses presented in Figures 2 and 3 to the individual WDS heuristic subscores. Figure 4 shows the marginal distribution of each score for the full corpus and for documents classified as low quality. These distributions describe the frequency of low-quality documents across the score range and should not be interpreted as the probability of a document being low quality given its score. Figure 5, in contrast, shows the proportion of documents classified as low quality within each score bin, obtained by normalising the number of low-quality documents by the total number of documents in the corresponding bin.

Score	Pearson's r	N
WDS score	-0.2790	127,957,098
lang_score	-0.0853	127,957,098
url_score	0.0024	127,957,098
punctuation_score	-0.4186	127,957,098
singular_chars_score	0.0143	127,957,098
numbers_score	-0.0444	127,957,098
repeated_score	-0.2299	127,957,098
informativeness_score	0.1424	127,957,098
n_long_segment_score	0.1002	127,957,098
great_segment_score	-0.1405	127,957,098

Table 3: Pearson correlation between individual quality scores and the target variable.

Corpora	**n-token/word**	**n-document/pages**
HPLT3 Persian	157.76B (t)	127.96M (d)
HPLT3-PerRef	105.9B (t)	102.37M (d)
Matina	72.9B (t)	-
Hakim	58.8B (t)	-
mC4 Persian	52B (t)	54M (p)
Targoman	41B (t)	75M (d)
Madlad-400 Persian	29.2B (t)	58.1M (d)
naab	15B (w)	-
HmBlogs	6.8B (t)	20M (d)
OSCAR Persian	4.3B (t)	-
ParsBERT	-	4M (d)
MirasText	1.4B (t)	2.8M (d)

Table 4: Persian Corpora size comparison

A.2 Optuna Hyperparameters Optimization

Hyperparameter optimization was performed using Optuna. The optimization searched over the learning rate, weight decay, warmup ratio, per-device training batch size, and gradient accumulation steps. The learning rate was sampled between 5×10^{-6} and 3×10^{-5} on a logarithmic scale. Weight decay was sampled between 0 and

0.1, while the warmup ratio was sampled between 0 and 0.15. The per-device training batch size was selected from 4 or 8, and the gradient accumulation steps were selected from 4 or 8.

For each trial, the model was initialized from the pretrained model and trained for up to 10 epochs. Early stopping was applied with a patience of 3 epochs. The per-device evaluation batch size was fixed at 16. Evaluation and logging were performed after each epoch, and the best model checkpoint was retained based on recall for class 0 on the development set (eval_recall_0). The optimization objective was therefore to maximize class-0 recall on the development set.

The remaining training settings were kept fixed across trials. These included a random seed of 42 and bfloat16 (BF16) mixed-precision training. The complete hyperparameter configuration selected by Optuna for the final model was: learning rate = 7.79×10^{-6} , weight decay = 0.0508, warmup ratio = 0.00184, per-device training batch size = 8, and gradient accumulation steps = 8. The configuration reported above corresponds to the trial achieving the highest class-0 recall on the development set.

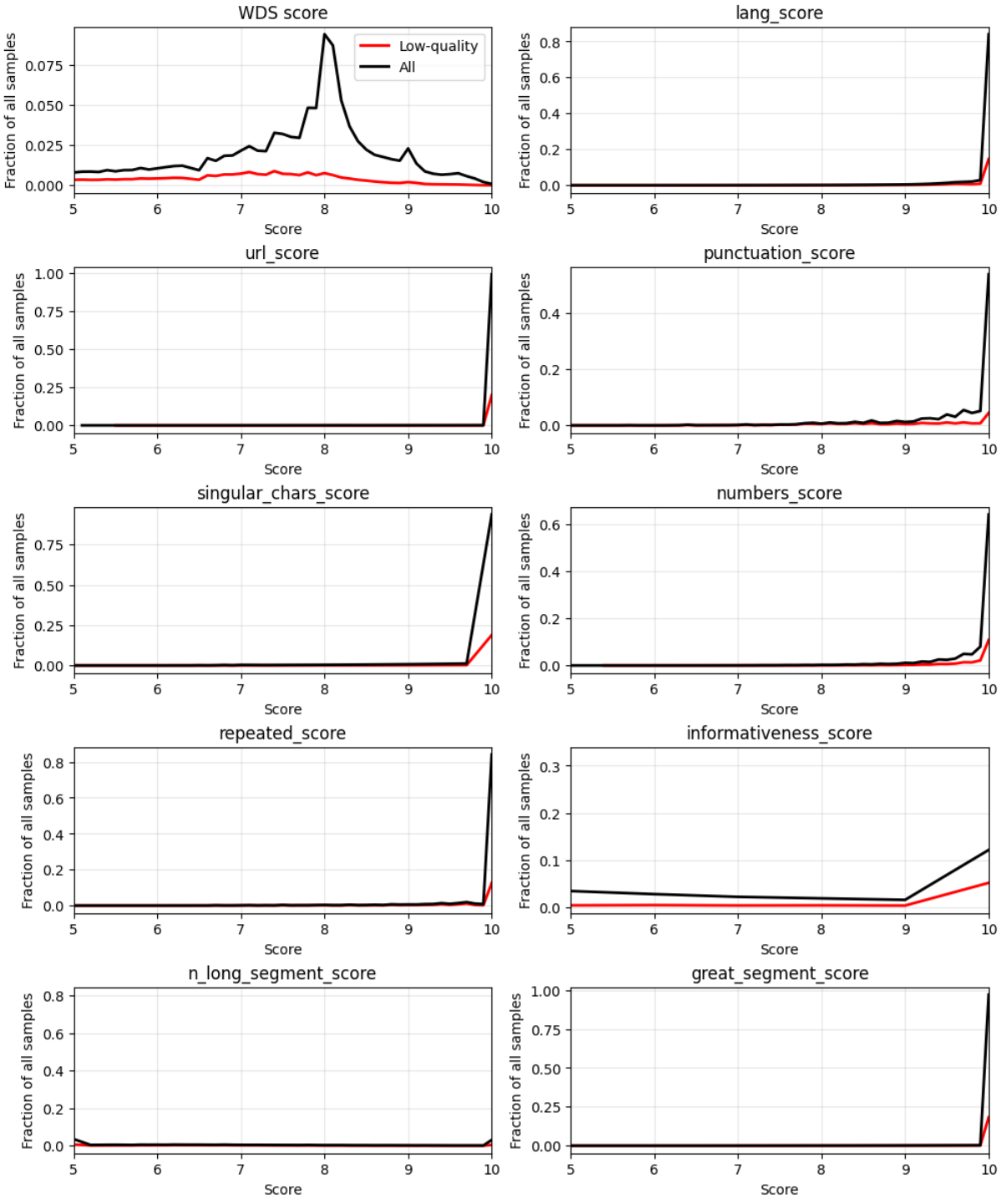


Figure 4: Marginal distributions of the WDS score and its heuristic subscores for the full HPLT3-PerRef corpus and documents classified as low quality. The WDS distribution corresponds to Figure 2 in the main text; the additional panels show the corresponding distributions for the individual heuristic subscores.

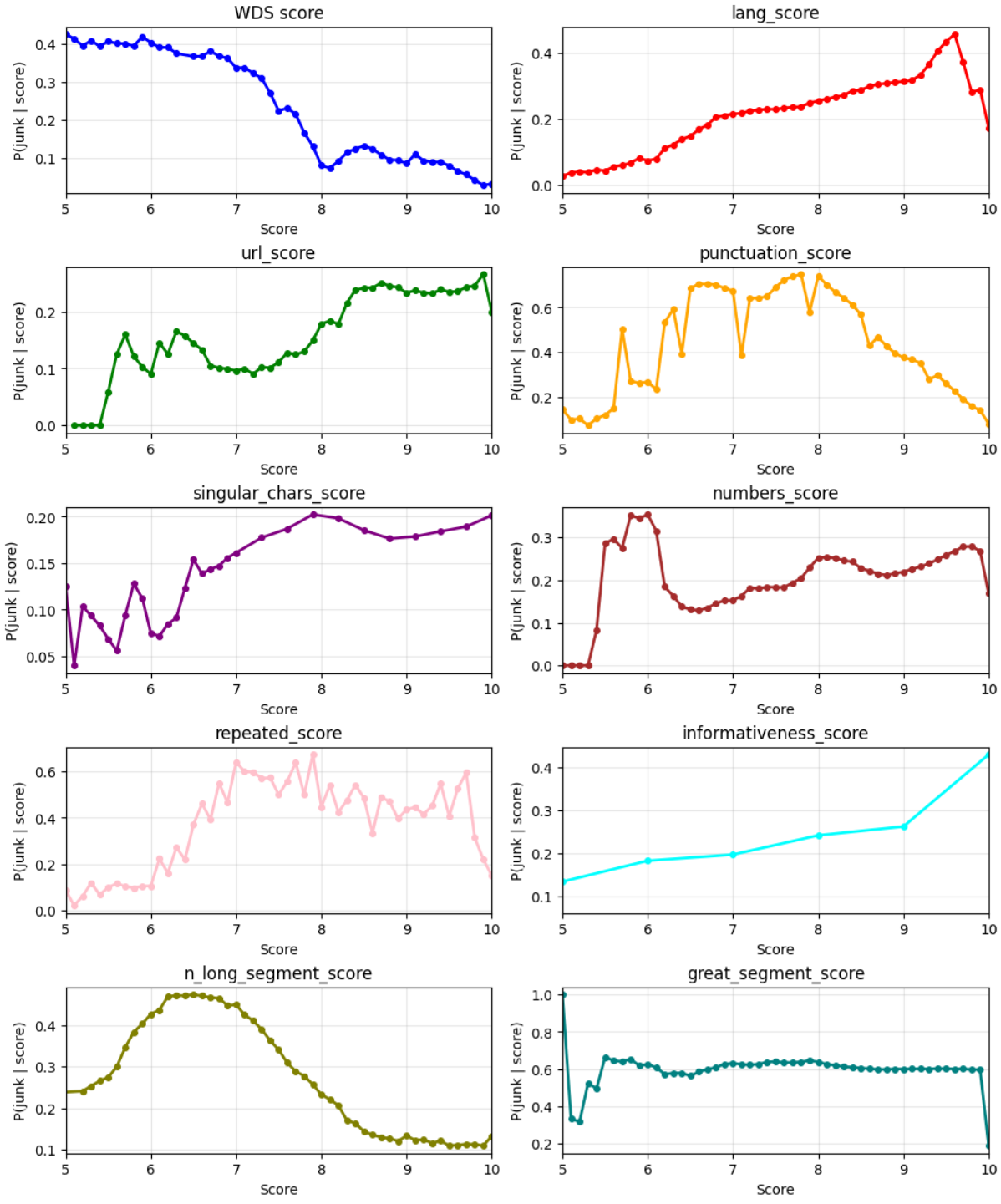


Figure 5: Proportion of documents classified as low quality within bins of the WDS score and its heuristic subscores. The WDS relationship corresponds to Figure 3 in the main text; the additional panels show the corresponding relationships for the individual heuristic subscores.

Host-Aware Web Crawling for LLM Pretraining Data

Malte Ostendorff
Common Crawl Foundation
malte@commoncrawl.org

Abstract

Web crawling is a major source of large language model (LLM) pretraining data, yet data-selection pipelines discard a large share of crawled pages. We investigate how a crawler should allocate its budget across and within hosts to collect more high-quality pretraining data. Combining document-level quality classifier scores, web-graph centrality metrics, and Common Crawl's per-host crawl coverage, we find that roughly half of the variation in classifier scores occurs between hosts. In the crawls we analyze, a few pages from a host also provide a useful estimate of the quality of its remaining pages. Common Crawl uses harmonic centrality, a graph centrality metric, to guide its crawl allocation. In our simulations, using centrality to choose which hosts to crawl yields limited gains over random ordering and performs worse for some classifiers. For deciding how deeply to crawl each host, sampling a few pages and then fetching more from hosts with higher-scoring samples collects several times more high-quality pages per fetch than uniform depth or centrality-based allocation. Finally, Common Crawl's observed per-host coverage has little correlation with the evaluated quality signals. We estimate that tens of millions of high-quality pages remain among known but uncrawled URLs in each crawl examined, with most concentrated on a small share of hosts.

1 Introduction

Large language model (LLM) pretraining relies heavily on Web data such as the Common Crawl (CC) corpus, which is collected by crawling. A Web crawler with a finite budget, like Common Crawl's crawler, must decide continuously *which* URLs to fetch next (Cho et al., 1998). CC decides this primarily based on a Web graph metric, namely harmonic centrality (HC; Boldi and Vigna, 2014).

Because most crawled pages are later discarded as low quality by pretraining pipelines (Li et al., 2024; Albalak et al., 2024), the crawl policy decision can be seen as a pretraining data selection or quality decision. However, the crawling aspect of pretraining data has been mostly neglected in the literature. Only recently, Yu et al. (2025) proposed with *Craw4LLM* a policy that prioritizes the crawl frontier by the pretraining-classifier score of each fetched page, reaching parity with a full crawl from a fraction of the URLs.

Such page-level schedulers are host-blind: they maintain a global per-page priority queue and must fetch, extract, and classify *every* candidate page to obtain the scheduling signal. We ask a complementary question at the host level, one that requires far less compute: given the Web graph and data-selection classifiers, how should a crawler allocate its budget across and within hosts - and is Common Crawl's deployed allocation already aligned with what pre-training quality selection would choose? To answer it, we study ten data-selection classifiers across several crawls and how they correlate with Web graph metrics and Common Crawl's own per-host crawl budget. The paper is organized around four research questions that lead to the host-level crawl policy.

(RQ1) Is document quality a host-level property? A host-level variance decomposition shows quality is moderately concentrated by host (ICC1 0.43–0.62), motivating our host-aware policy and providing the per-host sample size it needs (§4.1).

(RQ2) Do graph-centrality metrics work as a crawl-yield signal? A crawl-yield evaluation confirms and *explains* the negative graph-prior result of Yu et al. (2025) through a language confound and heterogeneous hub composition, across ten classifiers (§4.2).

(RQ3) How deep should a crawler go per host, and can a small sample decide? A host-aware crawl-*depth* analysis shows a ∼3-page sample pre-

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 70–87
October 29, 2026 ©2026 Association for Computational Linguistics

dicts a host's remaining quality, and a sample-then-deepen policy approaches the oracle policy while centrality-biased depth does not (§4.3).

(RQ4) Does Common Crawl's actual budget align with quality? An analysis of historical crawl data finds that CC's per-host budget is uncorrelated with quality under every classifier we test, leaving tens of millions of known high-quality URLs uncrawled on identifiable hosts (§4.4).

2 Related Work

LLM Pretraining Data. Pretraining datasets are typically filtered using methods that can process large volumes of text efficiently. These include hand-written rules based on document length, symbol frequency, stopword frequency, or language, as well as learned classifiers that estimate each document's quality from its text. The prevailing choice is the efficient fastText text classifier (Joulin et al., 2017), trained on high-quality reference corpora in English: DCLM (Li et al., 2024) uses a fastText OH-2.5+ELI5 classifier, GneissWeb (Gohari et al., 2025) an ensemble of five topic-specialized fastText classifiers, and Nemotron-CC (Su et al., 2024) a three-way fastText ensemble bucketed into five quality tiers; FineWeb-Edu (Penedo et al., 2024) uses a more compute-intensive Transformer model. The same recipe extends across languages: JQL-Edu (Ali et al., 2025) for over 35 languages, and the FineWeb2-HQ classifier (Messmer et al., 2025) for the multilingual FineWeb2 corpus (Penedo et al., 2025). Large open pretraining corpora such as Dolma (Soldaini et al., 2024) combine such signals into full, reproducible curation pipelines. All of these signals are content-internal, since they look only at the text and not at Web graph or host-level features. The only graph-adjacent signals in production pipelines are categorical domain blocklists and the social-graph seeds (Gokaslan and Cohen, 2019; Brack et al., 2024).

Graph Signals in Web IR. Classical link-analysis methods include PageRank (Page et al., 1999), HITS (Kleinberg, 1999), harmonic centrality (Boldi and Vigna, 2014), TrustRank (Gyöngyi et al., 2004), and anti-TrustRank. Castillo et al. (2007) showed that link-topology features classify Web spam better than content alone. CrediBench (Kondrup et al., 2025) trains a graph neural network on a 45M-node CC subgraph for source credibility prediction. LinkBERT (Yasunaga et al., 2022) mines QA pairs from Wikipedia hyperlink graphs for synthetic pretraining data. Both use graph structure for purposes other than document filtering.

URL-level Crawl Policy. *Craw4LLM* (Yu et al., 2025) is the closest prior work. On ClueWeb22 (Overwijk et al., 2022), the authors measure a weak Spearman $\rho \approx -0.11$ between a page's indegree and its DCLM score (on the ClueWeb22-B sample), and simulate crawling on the 900M-page ClueWeb22-A subset. Indegree-prioritised crawling reaches only 69.5% of oracle downstream LM performance, vs. 78.1% for random crawling and 95.3% for fasttext-driven crawling. They conclude that graph centrality is a poor standalone LLM data selector. Our results reproduce the negative overall correlation ($\rho = -0.119$ for HC on CC-MAIN-2024-10) and provide a plausible mechanism: English-trained classifier scores covary with host language composition, which itself covaries with HC. We do not analyse not-public ClueWeb22-A directly. Furthermore, we also find that PR is positively correlated with quality when comparing hosts within the same top-level domain (TLD).

Graph-based Data Selection. *WebGraphMix* (Badoni et al., 2026) evaluates mixtures of already-collected documents for LLM pretraining. We evaluate host ordering and depth allocation during crawling, measuring high-quality pages collected under a fixed fetch budget. We focus on harmonic centrality (HC) and PageRank (PR), the metrics Common Crawl publishes. WebGraphMix's finding that PR-based selection does not improve downstream performance is consistent with the limited crawl-yield gains from centrality-based ordering in our experiments (§4.2), although the two studies evaluate different setups. DCLM (Li et al., 2024) grouped documents into five bins by PR and found that no bin outperformed random sampling. Our results suggest a possible explanation: pooling hosts across TLDs can obscure the positive correlation between PR and quality within individual TLDs.

3 Data and Methods

Our experiments draw on *two distinct data regimes*, and keeping them separate is important for interpreting the results. In the **classified-sample** regime we do our own extraction and classification: we read WARC records, extract main text, and run our data-selection classifiers on documents extracted from 224 WARC files, approximately 0.25% of the

WARC files in CC-MAIN-2024-10 (RQ1–RQ4).[1] In the **full-crawl** regime, we instead use precomputed artifacts (Common Crawl's host index and the GneissWeb quality annotations covering *every* fetched page) without re-classification (RQ3 replication, RQ4). Appendix D lists every dataset and model and which research question uses it.

3.1 Classifying Web Documents

Our primary classified sample is from CC-MAIN-2024-10 (February/March 2024). Specifically, we extract the text from HTML pages using Trafilatura (Barbaresi, 2021) from 224 randomly selected WARC files. Keeping hosts with at least three documents yields 3.01M documents on 532,628 hosts for our primary host-level analyses. These document counts refer to our sample of WARC files, not to all pages that Common Crawl has fetched from each host. The depth analysis uses stricter inclusion thresholds, described in §4.3.

We score each extracted document with ten fasttext classifiers and take host-median scores: DCLM (Li et al., 2024), FineWeb-Edu (Penedo et al., 2024), three Nemotron-CC classifiers (quality, educational, information) (Su et al., 2024), and five GneissWeb classifiers (a quality annotator plus science, education, technology, and medical topic scorers) (Gohari et al., 2025). Six of the ten classifiers target pretraining *quality* and four score *topic*. Throughout, we use *quality* to mean a classifier's predicted usefulness for pretraining, which is itself only a proxy for downstream benchmark performance and not intrinsic document quality. For FineWeb-Edu, we run a compute-efficient fastText re-implementation (Tsui and Nguyen, 2024) of the original Transformer-based classifier, at comparable accuracy.

For the language analysis (§4.2), we additionally run two multilingual Transformer classifiers, JQL-Edu (Ali et al., 2025) and the FineWeb2-HQ classifier (Messmer et al., 2025; Penedo et al., 2025), on a smaller core set of 462,494 documents derived from 16 WARC files. Exact model identifiers are in Table 10. These scores drive the variance decomposition (RQ1), the crawl-yield simulation (RQ2), and the primary depth simulation (RQ3). We run the same extraction-and-scoring pipeline on a second snapshot, CC-MAIN-2023-40 (32 WARC files), which we use for the language and cross-crawl robustness checks (Appendix D).

3.2 Web Graph

Graph features come from Common Crawl's public host-level hyperlinkgraph[2], using the aligned `cc-main-2023-24` release (built from crawls spanning September 2023 to February 2024) so that a single graph covers our document snapshots. For each host we take harmonic centrality (HC) and PageRank (PR), the two centrality measures Common Crawl publishes, in $\log_{10}(1+x)$ form. Fewer than 0.3% of our hosts are missing from the graph and are excluded.

3.3 Crawl Budget and Quality Density

The full-crawl regime uses two Common Crawl artifacts we do *not* recompute ourselves. The **host index** reports, per host, the number of known URLs (`nutch_numRecords`), how many were fetched (`nutch_fetched`), and the resulting *crawl budget or coverage* (`nutch_fetched_pct`). The **GneissWeb annotation set** (Gohari et al., 2025), released by Common Crawl, labels the pages a strong quality filter kept; every annotated page passed the filter, and the annotations cover *every* fetched page rather than our WARC sample. Joining the two on the same crawl gives a per-host *quality density*,

$$\text{quality density} = \frac{\#\text{GW-quality pages}}{\#\text{fetched pages}}, \qquad (1)$$

the fraction of a host's crawled pages that are high quality. Because it is full-crawl and independent of our own extraction and classifiers, it both replicates the depth analysis at full scale (RQ3) and grounds the crawl-budget analysis (RQ4). We validate it against our own classifiers in §4.4. Concretely, the crawl-budget analysis joins the host index with GneissWeb quality annotations on CC-MAIN-2024-10 and also across six crawls for the temporal panel (CC-MAIN-{2021-49, 2022-49, 2023-23, 2023-40, 2024-10, 2024-18}). The full-crawl depth replication uses the GneissWeb depth annotations for CC-MAIN-2017-34.

3.4 Metrics

We use different metrics for different research questions. RQ1 examines variation in individual page scores. The graph-quality correlations use the median classifier score of each host. The crawl-yield

[1] Our own classification is needed since pretraining datasets are typically published without low quality samples and without the corresponding full range of classifier scores.

[2] https://commoncrawl.org/web-graphs

and depth simulations instead count pages whose classifier score exceeds a threshold and measure the proportion of high-quality pages. For RQ4, quality density is the proportion of fetched pages retained by the GneissWeb filter. A host's median score and its proportion of high-quality pages are different quantities.

Spearman ρ. Our primary statistic is Spearman rank correlation, which measures whether two quantities tend to rise together ($+1$) or fall as the other rises (-1), with 0 meaning no monotone relationship.

Host-level variance (RQ1). The intraclass correlation (Shrout–Fleiss one-way estimator)

$$\mathrm{ICC1} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}} \tag{2}$$

is the share of per-document score variance that lies *between* hosts. $\mathrm{ICC1} = 1$ would mean a host's pages are uniform in that score and 0 that a host tells us nothing about a page. With $\bar{m}$ documents per host, the Spearman–Brown reliability

$$\frac{\bar{m}\,\mathrm{ICC1}}{1 + (\bar{m} - 1)\,\mathrm{ICC1}} \tag{3}$$

bounds the host-level correlation any signal can achieve, so ICC1 also sets the per-host sample size needed to rank hosts.

Crawl-yield simulation (RQ2). We fix a crawl budget as a fraction of corpus documents and fetch hosts in the order a policy defines until the budget is spent. The policies are random, Common Crawl's current allocation, HC, PR, or an oracle that knows each host's true quality density. We report yield as the share of fetched pages that are high quality (Figure 2) and, for cross-classifier comparison, as a multiple of the random baseline (Appendix B).

Depth simulation (RQ3). We count a page as *high quality* when its quality score exceeds 0.5, and a host's quality density is its fraction of such pages, using both our FineWeb-Edu scores on the classified sample and GneissWeb quality scores on the full crawl. RQ1 already implies that a small per-host sample characterizes a host, but we also confirm this directly by correlating a k-page sample's quality density with that of the host's remaining pages (Figure 4), and then use the sample to decide crawl depth.

Each policy gets the same budget: a fixed number of pages to fetch, set as a fraction of all available pages. We score a policy by its *yield*, i.e., the high-quality pages it collects per page fetched. A crawler cannot tell a host's high quality pages from its low quality ones until it fetches them. So the expected high-quality count is the host's quality density times the number of pages fetched. A policy therefore cannot cherry-pick high quality pages *within* a host. Instead, it can only choose *which* hosts to crawl deeply. We compare four policies.

Uniform breadth crawls every host to the same shallow depth. HC *depth* crawls hosts fully in order of harmonic centrality, highest first, until the budget is spent. *Sample-then-deepen* first probes three pages per host, then crawls fully the hosts whose sampled quality density is the highest. The *oracle* crawls fully the hosts with the highest *true* quality density first, giving the upper bound.

Crawl-budget analysis (RQ4). We correlate each host's crawl coverage with its quality density, and define *hidden gems* as hosts below the median coverage whose quality density is in the top decile. We estimate the number of high-quality pages among a host's known but uncrawled URLs by multiplying their count by its observed quality density. This assumes that uncrawled pages are as likely to be high quality as the fetched pages from the same host.

4 Results

4.1 RQ1: Quality Is a Host-Level Property

Allocating more crawl budget to some hosts can improve yield if their pages tend to be higher quality than those on other hosts. We therefore measure how much of the variation in classifier scores occurs between hosts. Across all 10 classifiers, the between-host share of per-document score variance (ICC1) lies in 0.43–0.62, indicating moderate host-level consistency (Figure 1). In other words, roughly half of the observed variation in classifier scores is between hosts.

Moderate ICC1 has two consequences we use later. With $\bar{m} \approx 5$ observed documents per host in our WARC sample and $\mathrm{ICC1} \approx 0.5$, Spearman–Brown reliability is ≈ 0.83, so the largest host-level correlation any signal can reach is ≈ 0.9, not 1; near-zero correlations must be read against this ceiling. The observed host-level consistency also motivates testing whether a *small* sample predicts

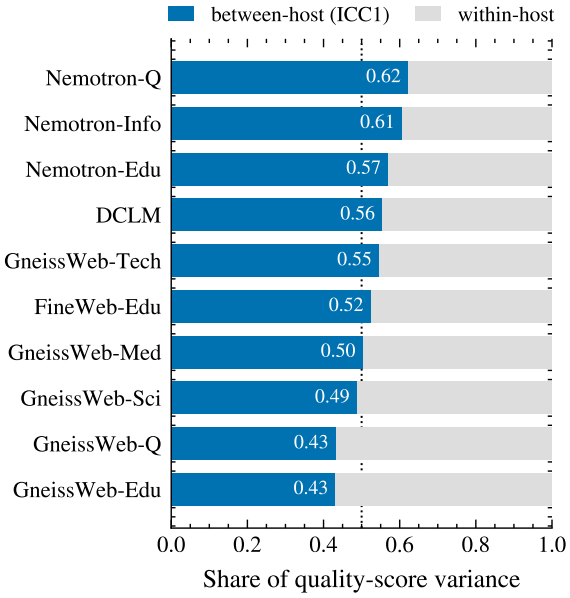


Figure 1: The intraclass correlation ICC1 is the share of a classifier's per-document score variance that lies *between* hosts rather than within them: Across all ten classifiers ICC1 is 0.43–0.62 ($N = 532{,}628$ hosts), so roughly half of the variation in page quality is between hosts. This suggests that a few sampled pages already characterize a host, enabling a host-level crawl policy.

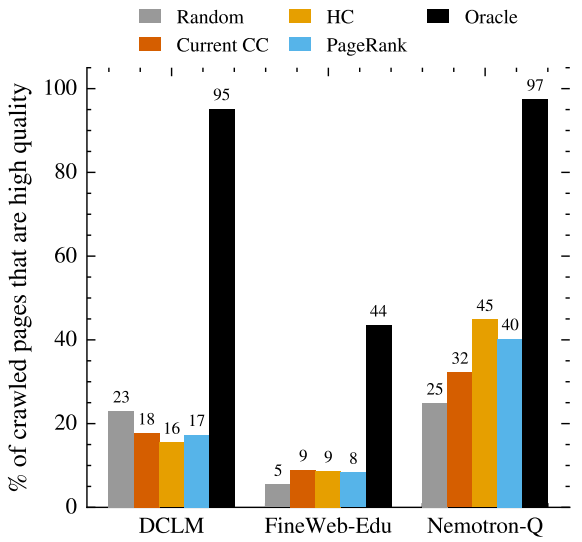


Figure 2: Graph centrality is a weak signal for *which* hosts to crawl. When only 10% of pages are crawled, the share of fetched pages that are high quality: centrality ordering (current/HC/PR) captures only a fraction of the *oracle* (crawler knows host's quality), and even trails random for DCLM.

the quality of remaining pages, which we evaluate directly in RQ3.

4.2 RQ2: Graph Centrality Is a Weak Crawl-Yield Signal

Common Crawl uses harmonic centrality (HC) to guide which hosts to crawl and how deeply to crawl them. We therefore test whether prioritizing hosts by centrality increases the number of high-quality pages collected under a fixed crawl budget.

At a 10% crawl budget, ordering hosts by HC collects 0.7–1.8 times as many high-quality pages as random ordering across classifiers, compared with 4.0–8.0 times for an oracle that knows each host's quality density (Figure 2). PageRank (PR) and Common Crawl's current allocation also yield substantially fewer high-quality pages than the oracle. For DCLM, centrality-based ordering performs worse than random. Ordering by HC also roughly halves the number of top-level domains represented compared with random ordering, reducing TLD diversity. Appendix B reports results by classifier and crawl budget.

Language partly explains the weak correlation. Prior work reports the similar findings. Craw4LLM measures a negative host-indegree/quality correla-

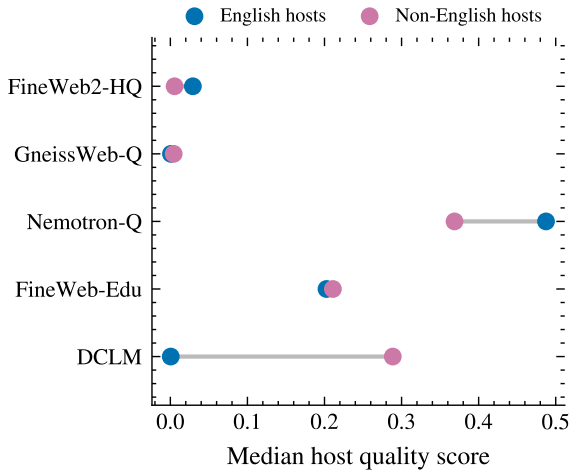


Figure 3: Median host quality scores by language. English-trained classifiers differ in how they score English and non-English hosts: DCLM assigns higher scores to non-English hosts, while Nemotron-Q assigns higher scores to English hosts. Because centrality also varies by language, its correlation with classifier scores partly reflects these language differences.

tion on ClueWeb22-A, and WebGraphMix rejects centrality for post-hoc selection. We reproduce the finding across ten classifiers on CC data and find that language differences partly explain the observed patterns.

A host's primary language covaries with its centrality, and the English-trained quality classifiers score English and non-English hosts very differently (Lucy et al., 2024), but not in a single direction (Figure 3). DCLM is the extreme case since it rates non-English hosts far higher than English ones (median 0.29 vs. 0.0005), so pooling the two languages makes its overall HC correlation *negative* ($\rho = -0.083$) even though it is positive within English hosts ($\rho = +0.064$). Nemotron-Q leans the opposite way, scoring English hosts higher (0.49 vs. 0.37). Multilingual classifiers, which do not share the bias, already show positive signal on non-English hosts (JQL-Edu +0.137, FineWeb2-HQ +0.186; full language-stratified table in Appendix C). Centrality is thus entangled with language and not a clean quality signal.

4.3 RQ3: A Small Per-Host Sample Decides Crawl Depth

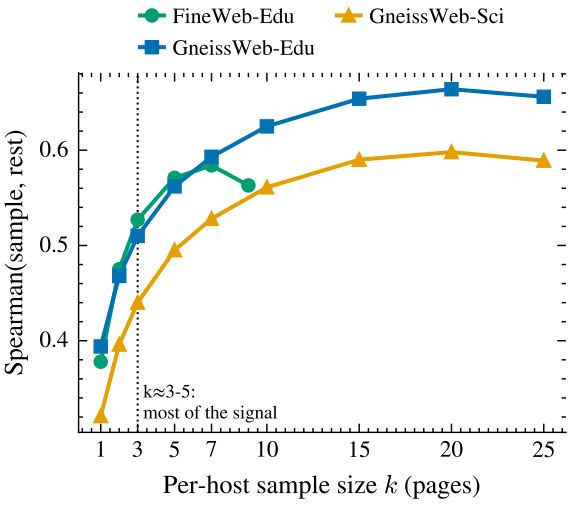


Figure 4: The correlation between the quality of k sampled pages and the quality of the host's remaining pages, where 1.0 is perfect prediction and 0 is none. The curve climbs with k and plateaus by $k \approx 15$–20, but a ∼3–5 page sample already captures most of the signal, so a few pages suffice to decide crawl depth cheaply. Shown for our FineWeb-Edu sample (hosts with ≥ 10 pages, so k reaches 9) and full-crawl GneissWeb (≥ 30 pages).

We now ask *how deep* to crawl each host. Because quality is a host property (RQ1), a few pages sampled from a host predict the rest: Figure 4 shows the rank correlation between a k-page sample and a host's remaining pages rises with k and plateaus by $k \approx 15$–20. The gain per added page shrinks steadily, so a small 3–5 page sample ($\rho = 0.53$ at k=3, 0.57 at k=5) already carries most of the signal. Ranking hosts on a three-page sample and deepening the best hosts already recovers most of the oracle's depth yield. A larger sample would rank hosts better at more computational costs.

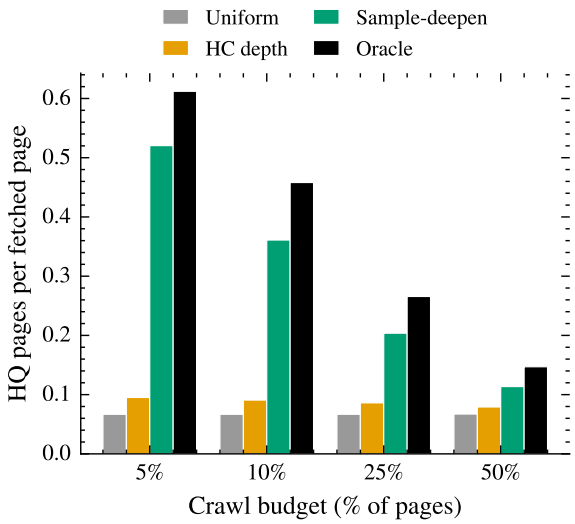


Figure 5: Crawl efficiency based on FineWeb-Edu, measured as high-quality pages obtained per page actually fetched, and for four ways of spending a fixed budget across hosts. *Sample-then-deepen* (our host-level crawl policy) nearly matches the *oracle* and collects several times more high-quality pages per fetch than crawling every host equally (*uniform breadth*) or biasing depth toward high-centrality hosts (HC *depth*).

This motivates a *sample-then-deepen* crawl policy, which samples and classifies a few pages per host, then spends depth on the hosts whose sample are higher quality. Such a policy obtains up to 7.8× more high-quality pages per fetched page than the uniform breadth policy at small budgets (down to 1.7× as the budget grows) and approaches the oracle policy.

In contrast, depth biased by harmonic centrality (similar to what CC does today) barely differs from uniform breadth (Figure 5). The *sample-then-deepen* policy has also the pratical advantage that the classifier is not in the crawler's hot path: unlike a page-level scorer (e.g., Craw4LLM) it runs on a ∼3-page sample per host, not on every page.

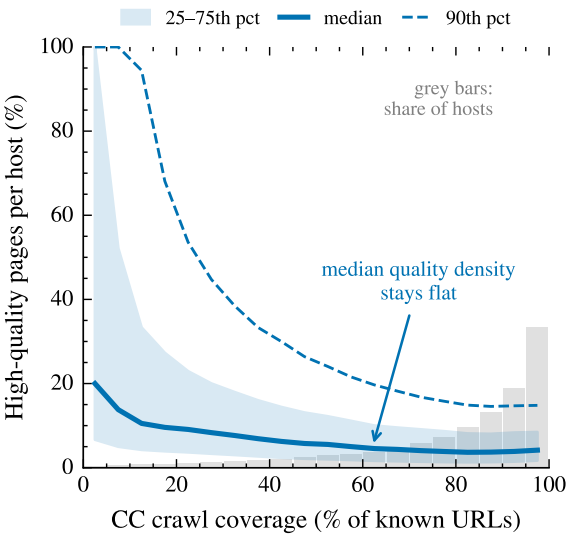


Figure 6: Host-level quality density does not rise with crawl coverage (Spearman -0.08). The median quality density stays flat and low across coverage (from 20% at the least-crawled hosts to 4% at the near-fully-crawled majority). Grey bars show the share of hosts in each coverage bin: most are already almost fully crawled, which is why the overall rank correlation is near zero.

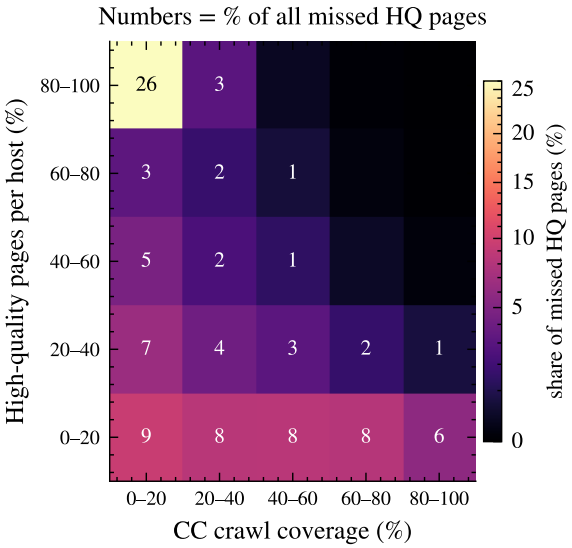


Figure 7: The missed high-quality budget concentrates on barely-crawled hosts. Each cell gives the share of the estimated ≈ 58M uncrawled high-quality pages that falls in one coverage×quality bin (columns: crawl coverage; rows: quality density; axes as in Figure 6). Half (50%) of the missed pages lie on hosts crawled below 20%, and the single densest "hidden-gem" bin holds 26% of missed high-quality pages alone.

4.4 RQ4: Common Crawl's Budget Is Blind to LLM Data Quality

Finally, we turn from simulation to the deployed crawl. Using historical per-host budget from CC-MAIN-2024-10 and GneissWeb quality scores, we find its crawl coverage carries no positive quality signal: across 5,654,186 hosts, the correlation between how fully a host is crawled and its quality density is -0.08, and the quality distribution does not rise with coverage (Figure 6). If anything the least-crawled hosts are the highest quality. Assuming that uncrawled pages have the same quality density as fetched pages from the same host, we estimate 58M high-quality pages among the known URLs left uncrawled in this one crawl. This missed budget concentrates on barely-crawled hosts: 50% of it lies on hosts crawled below 20%, and 62% falls on a compact set of 369,534 "hidden-gem" hosts (below-median coverage but top-decile quality) that represent only 6.5% of all hosts (Figure 7). The URLs are known, but the quality of their unfetched pages has not been directly measured.

Expressed as pretraining tokens, this single-crawl shortfall is on the order of tens of billions of high-quality tokens (roughly 58 billion at an assumed ∼1,000 tokens per page). We find shortfalls of a similar scale in each crawl examined (Table 1). Extrapolating this estimate to the +100 existing CC crawls would give approximately 6 trillion tokens before accounting for repeated pages or duplicate content. This is an illustrative extrapolation. We have not measured overlap across crawls or established that the same shortfall applies throughout the whole crawl archive.

Crawl	ρ(cov, qual)	Gems	Uncrawled HQ (M)
2021-49	−0.151	339,430	20.3
2022-49	−0.057	337,224	28.9
2023-23	−0.060	325,545	20.1
2023-40	−0.079	390,185	38.0
2024-10	−0.081	369,534	35.7
2024-18	−0.061	353,728	24.1

Table 1: The coverage–quality disconnect is not a one-off. For each of 6 crawls (2021–2024) we report the Spearman correlation between crawl coverage and quality density (all near 0 and slightly negative, i.e. no positive alignment), the number of under-crawled high-quality "gem" hosts, and the estimated high-quality URLs those hosts leave uncrawled. Every crawl shows the same pattern and tens of millions of missed URLs.

Stable and classifier-robust. The pattern is not a one-crawl or one-classifier artifact. Across six crawls spanning 2021–2024 the coverage–quality correlation stays near zero and negative (-0.057 to -0.151; Table 1). Also, the GneissWeb quality signal is confirmed by independent classifiers: Nemotron

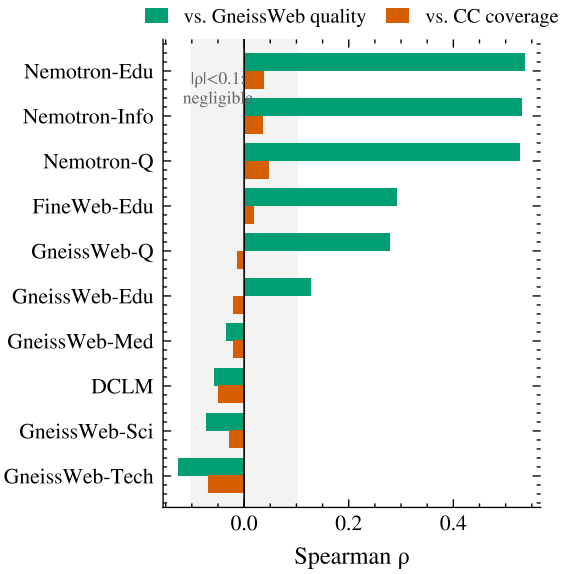


Figure 8: Coverage-blindness does not depend on the GneissWeb filter. Per classifier, on one Spearman axis: its agreement with the GneissWeb quality density (green) versus the correlation of Common Crawl coverage with it (orange). Quality classifiers (Nemotron, FineWeb-Edu) agree substantially with GneissWeb (DCLM and the topic scorers less so), yet every coverage bar stays inside the negligible band ($|\rho| < 0.1$): whichever quality signal we use, how fully a host is crawled does not track how good it is.

(0.53–0.54) and FineWeb-Edu (+0.29) agree with it, while Common Crawl coverage is uncorrelated with *every* classifier we test ($|\rho| \leq 0.07$; Figure 8). DCLM is the lone non-agreeing classifier (-0.06), consistent with its known orthogonality to other quality signals (§4.2). The missed budget skews toward national and regional news publishers and scholarly repositories: the .uk TLD alone accounts for 23% of it from 6.1% of hosts, at just 34% median coverage (Appendix, Figure 11).

5 Discussion

A host-level companion to Craw4LLM. Craw4LLM (Yu et al., 2025) prioritizes the crawl frontier by the quality score of each fetched page. Thus, a crawler must fetch, extract, and score every candidate to obtain that signal. Our results address a different, complementary lever. Because quality is a host property (RQ1) and a ∼3-page sample predicts a host's rest (RQ3), the scheduling signal for *how deep* to crawl a host can be obtained from a small sample rather than from every page, and the classifier need not live in the crawler's hot path. We answer "how much to crawl each host, and which under-crawled hosts are worth more," which the page-level, host-blind formulation does not address. Our analysis complements page-level scheduling and informs how much crawl budget should be allocate to each host.

Explaining earlier findings on graph centrality. Craw4LLM and WebGraphMix (Badoni et al., 2026) report limited benefits from the centrality measures they evaluate. Our crawl-yield results show a similar pattern (RQ2): centrality-based ordering performs much worse than the oracle and, for DCLM, worse than random ordering.

Language differences help explain some of the observed centrality–quality correlations. English-trained classifiers score English and non-English hosts differently. For instance, DCLM assigns higher scores to non-English hosts. Centrality also varies by language. The correlation between HC and DCLM scores is negative when all hosts are pooled, but positive when only English-language hosts are considered. High-centrality hosts also contain a mix of high- and low-quality pages, limiting centrality's usefulness as a quality signal.

However, the language analysis does not establish that accounting for language improves crawl yield. A positive correlation within a language group need not translate into an effective ranking under a limited crawl budget. Likewise, positive correlations with multilingual classifiers do not by themselves demonstrate a better crawl policy. Lastly, centrality could still be used to group hosts for sampling, although we do not evaluate whether this improves crawl yield.

What the deployed crawl leaves on the table. RQ4 turns the diagnostic on CC's deployed crawl policy itself. Its per-host budget is uncorrelated with quality under every classifier we tested, so the crawl is not preferentially deepening the hosts worth deepening. This misalignment is expected rather than a defect: Common Crawl was built as a broad, general-purpose sample of the Web serving diverse use cases, not as an LLM-oriented crawler (Baack, 2024).

Assuming that uncrawled pages have the same quality density as fetched pages from the same host, we estimate that tens of millions of high-quality pages remain among known but uncrawled URLs in each crawl. The presented *sample-then-deepen* approach could help recover these pages by allocating more budget to hosts with high-scoring samples. It is worth noting that this gap is about *known*

URLs, for an estimation of whole URL population we refer to Paris et al. (2026).

Quality and discovery are different objectives. Our depth simulations allocate a crawl budget over a fixed set of already-fetched pages. They do not measure which additional URLs or hosts a crawler would discover by following links. A host with low classifier scores could nevertheless provide links to useful content elsewhere. Weak performance as a quality-ranking signal therefore does not rule out a role for centrality in discovery. A crawler could combine centrality-based exploration with sample-guided depth allocation, but our experiments do not evaluate this combination.

6 Conclusions

In this paper, we studied Web crawl policies for LLM pretraining data at the level of hosts and reached the following conclusions.

Document quality is largely a host property, i.e, between-host variance is about half of the total. This finding motivates the host-aware policy and lets a small sample stand in for a host. Ordering hosts by graph centrality (HC or PR) collects at most 1.8 times as many high-quality pages as random ordering, and fewer for DCLM. An oracle policy that knows each host's quality density collects 4.0–8.0 times as many. Centrality is a weak guide partly because it varies with host language, which also affects classifier scores, and because highly central hosts contain a mix of high- and low-quality pages.

In our crawl simulations, a three-page sample provides a useful signal for allocating crawl depth. The sample-then-deepen policy approaches the oracle high-quality yield, reaching up to 7.8 times the yield of the uniform breadth policy, whereas centrality-biased depth provides little to none improvement. Finally, Common Crawl's deployed crawl budget is itself blind to LLM data quality. The crawl coverage is uncorrelated with quality (Spearman -0.08) under every classifier, leaving an estimated 58M known high-quality URLs uncrawled in a single crawl, 62% of them on a compact set of 369,534 under-crawled hosts. This finding can be attributed to Common Crawl aiming for a broad, general-purpose sample of Web, and not LLM pretraining data.

In summary, our results motivate a crawl policy that first samples a few pages from each host, scores their quality, and then fetches more pages from hosts with higher-scoring samples. This could recover high-quality pages missed by the current crawl policy. Graph centrality could help group hosts for sampling, but should not be used on its own to judge their content quality.

6.1 Future Work

A live sample-then-deepen crawl. All four research questions rely on existing data from Common Crawl and did not involve any live crawling. RQ1 measures host-level variation in classifier scores, while RQ4 examines observed crawl coverage and quality density. RQ2 and RQ3 additionally simulate alternative crawl allocations over already-fetched pages. None of these analyses tests a newly deployed crawler under production settings, future work needs to address this. In particular, a live crawl experiment must account for fetch latency, content freshness, politeness limits, and the discovery of new URLs and hosts.

Recovering the hidden gems. RQ4 identifies tens of millions of known high-quality URLs on under-crawled hosts. A targeted re-crawl of the top gem hosts, followed by a downstream ablation, would quantify how much the missed budget is worth in trained-model terms.

Downstream model ablation. We measure crawl yield and coverage, not trained-model performance. Given that proxy metrics and downstream accuracy can diverge (Li et al., 2024; Yu et al., 2025), one would need to train small models ($\sim$1B parameters) on data selected by uniform, centrality, and sample-then-deepen policies test whether the yield gains translate into downstream quality.

Beyond a single quality filter. The quality-density signal uses primarily a single filter (GneissWeb) due to the availability of precomputed annotations. Combining several classifiers, and modeling their per-host agreement, may sharpen both the depth decision and the hidden-gem ranking.

Limitations

The following discussion complements the limitations described in §6.1.

Quality density is a proxy. Our per-host quality signal is the fraction of a host's crawled pages retained by a classifier model (GneissWeb). It inherits that models's biases. We mitigate this by showing the RQ4 result holds under independent models (Nemotron, FineWeb-Edu), but no model is ground truth (Gururangan et al., 2022).

Missed-budget estimate. The uncrawled-high-quality-URL count multiplies a host's uncrawled known URLs by its observed quality density, assuming the uncrawled pages resemble the crawled ones. It is an estimate of scale, not an exact count. Furthermore, the estimate does not account for overlap between crawls and other factors like crawl politeness or content freshness.

Selection of pages and hosts. Both the host-consistency estimates and the sample-size analysis concern pages already selected by Common Crawl. Our WARC sample introduces a further sampling step, and minimum-document thresholds exclude sparsely represented hosts. The full-crawl GneissWeb analysis removes the WARC subsampling step, but still reflects Common Crawl's page selection. These choices may affect measured within-host variation and how well a small sample predicts other pages. Our results therefore do not establish that three pages suffice for arbitrary hosts or for pages outside the observed crawls.

Representation and diversity. Allocating more crawl budget to hosts with high classifier scores may reduce the representation of languages, regions, or content types that those classifiers score poorly. At crawl time, this affects which pages are collected at all, including potentially useful pages on hosts whose initial samples receive low scores. Higher classifier-based yield therefore does not necessarily imply a more representative or diverse corpus. A deployment could reserve budget for broader exploration and monitor coverage across languages and content types, but the effectiveness of these measures remains to be evaluated.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research (TMLR)*. ArXiv:2402.16827.

Mehdi Ali, Manuel Brack, Max Lübbering, Elias Wendt, Abbas Goher Khan, Richard Rutmann, Alex Jude, Maurice Kraus, Alexander Arno Weber, David Kaczér, Florian Mai, Lucie Flek, Rafet Sifa, Nicolas Flores-Herr, Joachim Köhler, Patrick Schramowski, Michael Fromm, and Kristian Kersting. 2025. Judging quality across languages: A multilingual approach to pretraining data filtering with language models. ArXiv:2505.22232.

Stefan Baack. 2024. A critical analysis of the largest source for generative AI training data: Common Crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Vedant Badoni, Danqi Chen, and Xinyi Wang. 2026. Hubs or fringes: Pretraining data selection via web graph centrality. ArXiv:2606.11499.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference 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. Association for Computational Linguistics.

Paolo Boldi and Sebastiano Vigna. 2014. Axioms for centrality. *Internet Mathematics*, 10(3–4):222–262.

Manuel Brack, Malte Ostendorff, Pedro Ortiz Suarez, José Javier Saiz, Iñaki Lacunza Castilla, Jorge Palomar-Giner, Alexander Shvets, Patrick Schramowski, Georg Rehm, Marta Villegas, and Kristian Kersting. 2024. Community OSCAR: A community effort for multilingual web data. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 232–235, Miami, Florida, USA. Association for Computational Linguistics.

Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, and Sebastiano Vigna. 2007. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th Annual International ACM SIGIR Conference*, pages 423–430.

Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. 1998. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172.

Hajar Emami Gohari, Swanand Ravindra Kadhe, Syed Yousaf Shah, Constantin Adam, Abdulhamid Adebayo, Praneet Adusumilli, Farhan Ahmed, Nathalie Baracaldo Angel, Santosh Subhashrao Borse, Yuan-Chi Chang, Xuan-Hong Dang, Nirmit Desai, Revital Eres, Ran Iwamoto, Alexei Karve, Yan Koyfman, Wei-Han Lee, Changchang Liu, Boris Lublinsky, and 13 others. 2025. GneissWeb: Preparing high quality data for LLMs at scale. Technical report, IBM Research. ArXiv:2502.14907.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText corpus.

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2201.10474.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 576–587.

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 (EACL)*, pages 427–431. ArXiv:1607.01759.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Emma Kondrup, Sebastian Sabry, Hussein Abdallah, Zachary Yang, Jiaqi Xiong, Kellin Pelrine, James Zhou, Zhijin Guo, Michael M. Bronstein, Jean-François Godbout, Reihaneh Rabbany, and Shenyang Huang. 2025. CrediBench: Building web-scale network datasets for information integrity. ArXiv:2509.23340.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*. ArXiv:2406.11794.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren F. Klein, and Jesse Dodge. 2024. AboutMe: Using self-descriptions in webpages to document the effects of English pretraining data filters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv:2401.06408.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. 2025. Enhancing multilingual LLM pretraining with model-based data selection. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*. ArXiv:2502.10361.

Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with visual and semantic information. *Preprint*, arXiv:2211.15848.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Stanford Technical Report, ILPUBS-422.

Michael Paris, Hande Celikkanat, and Luca Foppiano. 2026. Measuring what the crawler sees: Discovery curves, core persistence, and shell dynamics in longitudinal web crawls. *Preprint*, arXiv:2607.13636.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*. ArXiv:2406.17557.

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. Technical report, Hugging Face. ArXiv:2506.20920.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv:2402.00159.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. Technical report, NVIDIA. ArXiv:2412.02595.

Ken Tsui and Huu Nguyen. 2024. Low latency CPU based educational value classifier with generic educational value. Hugging Face model.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8003–8016. ArXiv:2203.15827.

Shi Yu, Zhiyuan Liu, and Chenyan Xiong. 2025. Craw4LLM: Efficient web crawling for LLM pretraining. In *Findings of the Association for Computational Linguistics: ACL 2025*. ArXiv:2502.13347.

A Temporal Stability of Graph Ranks

Later release	Gap	N hosts	HC ρ	PR ρ
2024-may-jun-jul	∼3-5mo after	531,054	0.875	0.923
2024-25-nov-dec-jan	∼9-11mo after	527,432	0.870	0.900
2025-may-jun-jul	∼15mo after	524,077	0.851	0.885

Table 2: Spearman rank correlation between a host's centrality rank in our reference graph and its rank in later Common Crawl graph releases, where 1.0 would mean an identical ordering of all hosts. Both harmonic centrality (HC) and PageRank (PR) stay above 0.85 even ∼15 months apart, so host rankings barely move over a year. This stability is what lets us pair one aligned graph with document snapshots from different crawls instead of rebuilding the graph each time.

We assess how host centrality rankings change over time by comparing the reference graph with releases up to approximately 15 months later. Host-level HC and PR ranks have Spearman correlations of 0.851–0.923 across these comparisons (Table 2). This stability supports using a single reference graph for the document snapshots in our study.

B Crawl-Yield Detail (RQ2)

Classifier	Current	HC	PR	Oracle	TLDs (HC/rnd)
DCLM	0.75×	0.67×	0.73×	**3.98**×	223 / 422
FineWeb-Edu	1.60×	1.58×	1.47×	**7.97**×	223 / 422
Nemotron Quality	1.27×	1.79×	1.56×	**4.05**×	223 / 422

Table 3: High-quality documents collected at a 10% crawl budget, as a multiple of a random crawl (1.0 =random, 2× =twice as many). Centrality ordering (HC, PR) gains at most 1.8×—and drops *below* 1 for DCLM—while the oracle reaches 4.0–8.0×, so most of the achievable gain is left untaken. The last column shows centrality ordering also roughly halves the number of top-level domains represented (HC vs. random), narrowing corpus diversity. This is the per-classifier, fixed-budget companion to Figure 2; the full budget sweep is Figure 9.

This section gives the per-classifier crawl-yield numbers behind §4.2. Table 3 fixes the budget at 10% of pages and reports, for each classifier, the high-quality yield of each ordering policy as a multiple of a random crawl, together with the TLD diversity the policy retains; Figure 9 plots the same comparison across all budgets. It makes the two headline points of RQ2 concrete at a single budget: centrality ordering barely lifts the yield above random (and drops below it for DCLM) while the oracle is several times higher, and centrality ordering simultaneously narrows the range of top-level domains crawled.

C Full Graph-Classifier Correlation Tables

These tables expand the correlation analysis behind RQ1–RQ2 using alternative classifiers, host-inclusion thresholds, and aggregation methods.

C.1 HC and PageRank Correlations Across Classifiers

Table 4 reports correlations of host-level HC and PR with all twelve classifiers, using Spearman ρ, Kendall τ, and weighted Spearman ρ. Spearman correlations range from -0.086 to $+0.207$. For each classifier, HC and PR have the same correlation sign and similar magnitudes.

Metric	Classifier	Spearman ρ	Kendall τ	Weighted ρ	N hosts
log HC	DCLM	−0.083	−0.055	−0.087	27,910
log HC	GneissWeb Quality	+0.021	+0.015	+0.018	27,910
log HC	GneissWeb Sci	+0.031	+0.021	+0.030	27,910
log HC	GneissWeb Edu	+0.039	+0.026	+0.037	27,910
log HC	GneissWeb Tech	+0.036	+0.024	+0.038	27,910
log HC	GneissWeb Med	−0.032	−0.021	−0.035	27,910
log HC	FineWeb-Edu	+0.122	+0.083	+0.125	27,910
log HC	Nemotron Quality	+0.167	+0.110	+0.169	27,910
log HC	Nemotron Edu	+0.156	+0.106	+0.154	27,910
log HC	Nemotron Info	+0.149	+0.100	+0.151	27,910
log HC	JQL-Edu	+0.124	+0.083	+0.120	15,661
log HC	FineWeb2-HQ	+0.207	+0.140	+0.217	13,614
log PR	DCLM	−0.086	−0.057	−0.089	27,910
log PR	GneissWeb Quality	+0.012	+0.008	+0.009	27,910
log PR	GneissWeb Sci	+0.015	+0.010	+0.015	27,910
log PR	GneissWeb Edu	+0.023	+0.015	+0.021	27,910
log PR	GneissWeb Tech	+0.029	+0.019	+0.032	27,910
log PR	GneissWeb Med	−0.039	−0.026	−0.042	27,910
log PR	FineWeb-Edu	+0.110	+0.075	+0.113	27,910
log PR	Nemotron Quality	+0.170	+0.113	+0.172	27,910
log PR	Nemotron Edu	+0.148	+0.101	+0.147	27,910
log PR	Nemotron Info	+0.152	+0.101	+0.153	27,910
log PR	JQL-Edu	+0.079	+0.053	+0.076	15,661
log PR	FineWeb2-HQ	+0.173	+0.117	+0.184	13,614

Table 4: Host-level correlations of harmonic centrality and PageRank with twelve classifier scores. Spearman ρ, Kendall τ, and weighted Spearman ρ agree in sign for every metric–classifier pair.

C.2 Language-Stratified Correlations (log **HC,** $n \geq 3$)

Table 5 is the full version of the language split behind Figure 3: the HC–quality correlation for every classifier, computed on all hosts and separately on English- and non-English-primary hosts. It shows the sign flip is general (classifiers tend to correlate more positively within a single language than pooled across languages) and that the multilingual classifiers (JQL-Edu, FineWeb2-HQ) are the ones that retain a positive signal on non-English hosts.

Stratum	Classifier	Spearman ρ	N hosts
English	FineWeb-Edu	+0.237	13,670
English	FineWeb2-HQ	+0.233	1,109
English	GneissWeb Sci	+0.199	13,670
English	Nemotron Edu	+0.157	13,670
English	GneissWeb Tech	+0.154	13,670
English	GneissWeb Edu	+0.143	13,670
English	Nemotron Quality	+0.127	13,670
English	GneissWeb Med	+0.111	13,670
English	Nemotron Info	+0.097	13,670
English	JQL-Edu	+0.083	1,421
English	GneissWeb Quality	+0.072	13,670
English	DCLM	+0.064	13,670
Non-English	FineWeb2-HQ	+0.186	12,505
Non-English	JQL-Edu	+0.137	14,186
Non-English	Nemotron Edu	+0.043	14,186
Non-English	FineWeb-Edu	+0.023	14,186
Non-English	GneissWeb Sci	+0.023	14,186
Non-English	GneissWeb Quality	+0.017	14,186
Non-English	DCLM	+0.017	14,186
Non-English	Nemotron Info	+0.016	14,186
Non-English	Nemotron Quality	+0.012	14,186
Non-English	GneissWeb Edu	+0.002	14,186
Non-English	GneissWeb Med	+0.001	14,186
Non-English	GneissWeb Tech	−0.007	14,186

Table 5: Full language-stratified HC correlations. For JQL-Edu and FineWeb2-HQ, "English" rows reflect host-primary language classification; inference excluded documents mapped as English.

C.3 TLD Macro-Category Breakdown

Table 6 breaks the HC–DCLM correlation down by TLD macro-category. It is negative precisely within the academic and commercial categories, which concentrate the non-English hosts that DCLM scores highly, i.e., localising the language confound of §4.2 to the domains where language and centrality covary most.

TLD Category	HC vs. DCLM ρ	N hosts
Other	−0.159	659
Academic (.edu, .ac.*)	−0.098	766
Commercial (.com, .net)	−0.028	13,470
Country-code	+0.004	10,939
Nonprofit (.org)	+0.032	1,881
Government (.gov.*)	+0.050	195

Table 6: HC–DCLM Spearman ρ by TLD macro-category. Negative within academic and commercial TLDs reflects DCLM's language inflation of non-English domains.

C.4 Score Distributions by Language Stratum

Table 7 gives the full score distributions (mean, median, and percentiles) behind Figure 3, split by host language. DCLM's median score is substantially higher on non-English hosts than on English hosts, whereas FineWeb-Edu assigns similar scores to the two groups. These differences are consistent with language composition contributing to the pooled centrality-quality correlation.

Stratum	Classifier	Mean	Median	P10	P90
English	DCLM	0.041	0.001	0.000	0.065
English	FineWeb-Edu	0.220	0.203	0.041	0.408
English	Nemotron Quality	0.541	0.487	0.396	0.785
English	JQL-Edu	-4.850	-5.409	-8.946	0.190
English	FineWeb2-HQ	0.242	0.029	0.000	0.958
Non-English	DCLM	0.367	0.289	0.002	0.892
Non-English	FineWeb-Edu	0.223	0.211	0.020	0.417
Non-English	Nemotron Quality	0.352	0.369	0.238	0.400
Non-English	JQL-Edu	-4.478	-4.897	-8.167	-0.096
Non-English	FineWeb2-HQ	0.132	0.005	0.000	0.586

Table 7: Score distributions by language stratum (selected classifiers). JQL-Edu scores are log-probabilities.

C.5 Full ICC1 Table

Table 8 lists the full RQ1 variance decomposition: ICC1 for every classifier at $n \geq 3$ documents per host. The between-host share is stable across the $n \geq 3$, 5, and 10 thresholds and near one-half throughout, the property that lets a small per-host sample stand in for a host (§4.1).

Classifier	n_{hosts}	n_{docs}	$\bar{m}$	σ^2_{btw}	σ^2_{wtn}	ICC1
Nemotron Quality	27,950	104,098	3.72	0.0195	0.0105	0.650
Nemotron Info	27,950	104,098	3.72	0.0379	0.0222	0.630
Nemotron Edu	27,950	104,098	3.72	0.0281	0.0190	0.596
GneissWeb Tech	27,950	104,098	3.72	0.0634	0.0461	0.579
DCLM	27,950	104,098	3.72	0.0596	0.0461	0.564
FineWeb-Edu	27,950	104,098	3.72	0.0131	0.0109	0.545
JQL-Edu	15,690	52,687	3.36	7.3071	6.2660	0.538
GneissWeb Med	27,950	104,098	3.72	0.0307	0.0275	0.528
GneissWeb Sci	27,950	104,098	3.72	0.0338	0.0319	0.514
FineWeb2-HQ	13,639	45,320	3.32	0.0485	0.0495	0.495
GneissWeb Edu	27,950	104,098	3.72	0.0153	0.0176	0.465
GneissWeb Quality	27,950	104,098	3.72	0.0154	0.0181	0.459

Table 8: Full ICC1 table ($n \geq 3$ docs per host). ICC1 stable across $n \geq 3/5/10$ thresholds.

D Datasets and Models

Our study combines several public Common Crawl artifacts and Hugging Face models across the two data regimes of §3. Table 9 lists every dataset and which research question uses it. Table 10 lists the exact classifier models we run in the classified-sample regime.

Dataset	Description and role (research question)
CC-MAIN-2024-10	Crawl (WARC) for primary crawl; we extract text (Trafilatura) and score it ourselves (224 WARCs, RQ1–RQ4).
CC-MAIN-2023-40	Crawl (WARC) for second snapshot and the language/robustness checks (32 WARCs, RQ2).
CC-MAIN-{2021-49, 2022-49, 2023-23, 2023-40, 2024-10, 2024-18}	Crawl (host stats from six crawls) for the temporal panel (RQ4).
Graph release (Sep 2023–Feb 2024)	Aligned host-level graph with harmonic centrality and PageRank per host (RQ2, App. A).
Later graph releases (2024-May, 2024-Nov, 2025-May)	Host HC/PR ranks at ∼3–15 months later, for rank stability (App. A).
CC host index	Per-host known/fetched URL counts and crawl coverage (RQ4).
GneissWeb annotations	Precomputed annotations for full-crawl page-level quality labels (every fetched page); source of the per-host quality density (RQ3, RQ4).

Table 9: Datasets used, by data regime; the research question each supports is noted in parentheses. WARC crawls are extracted and scored by us (classified-sample regime); the host index and GneissWeb annotations are Common Crawl's precomputed full-crawl artifacts.

Classifier	Family	Hugging Face model	Notes
DCLM	fastText	`mlfoundations/fasttext-oh-eli5`	OH-2.5 + ELI5 vs. RefinedWeb, bigram
GneissWeb Quality	fastText	`ibm-granite/GneissWeb.Quality_annotator`	quality annotator
GneissWeb Sci	fastText	`ibm-granite/GneissWeb.Sci_classifier`	science topic
GneissWeb Edu	fastText	`ibm-granite/GneissWeb.Edu_classifier`	education topic
GneissWeb Tech	fastText	`ibm-granite/GneissWeb.Tech_classifier`	technology topic
GneissWeb Med	fastText	`ibm-granite/GneissWeb.Med_classifier`	medical topic
FineWeb-Edu	fastText	`kenhktsui/fineweb-edu-fasttext-classifier`	educational value; fastText re-impl.
Nemotron (Q/Edu/Info)	fastText	`nvidia/nemotron-climb-fasttext-classifiers`	three heads: quality, edu, info
JQL-Edu	Transformer	`JQL-AI/JQL-Edu-Heads` on `Snowflake/snowflake-arctic-embed-m-v2.0`	multilingual (35 lang.); non-English only
FineWeb2-HQ	Transformer	`epfml/FineWeb-HQ-Classifiers` on `FacebookAI/xlm-roberta-base`	multilingual (21 lang.); non-English only

Table 10: Exact Hugging Face models used to classify pages in the classified-sample regime. FineWeb-Edu is a compute-efficient fastText re-implementation (Tsui and Nguyen, 2024) of the original Transformer classifier, at comparable accuracy. The Transformer classifiers (JQL-Edu, FineWeb2-HQ; the latter is the classifier of Messmer et al., 2025) are frozen-embedding models with lightweight per-language heads and are run on non-English documents only.

E Supplementary Figures

These figures are the visual companions to results summarised in the body: the full crawl-yield budget sweep behind Figure 2 (§4.2), the temporal robustness of the coverage–quality disconnect (Table 1), the composition of the missed high-quality budget (§4.4), and the agreement structure among the twelve classifiers (§4.2).

Crawl yield across budgets. Figure 9 is the full-budget companion to the fixed-budget bars of Figure 2. Sweeping the crawl budget from a small fraction of pages up to the full crawl shows that centrality ordering trails the oracle at *every* budget: the gap is widest at the small budgets where crawl order matters most and closes only as the crawl approaches completion, so the fixed-10% snapshot in the body is representative rather than a cherry-picked operating point.

Persistence across crawls. Figure 10 is the plotted form of Table 1: the coverage–quality disconnect of RQ4 holds in each of 6 crawls spanning 2021–2024, both as a near-zero, slightly negative correlation (panel a) and as tens of millions of uncrawled high-quality URLs per crawl (panel b). The misallocation is thus a standing property of the deployed crawl, not a single-crawl artifact.

Where the missed budget sits. Figure 11 characterises the hidden-gem hosts in aggregate (no individual host is named). Country-code publisher TLDs—led by .uk—are heavily over-represented relative to their host count, and among the hosts a keyword classifier can label, news/media outlets and scholarly repositories dominate. This is what the recoverable pool of §4.4 concretely consists of, and it motivates the targeted re-crawl proposed in the future work.

How the classifiers relate. Figure 12 gives the pairwise Spearman correlation between the twelve classifiers' scores across the 462,494 core documents (every document scored by all classifiers). It contextualises two threads in the body. DCLM is largely orthogonal to the other quality signals—weakly correlated with most and *negative* against the Nemotron heads (-0.459 with Nemotron-Q)—consistent with its outlier behaviour throughout (§4.2) and its lone non-agreement in the coverage-robustness check (Figure 8). The GneissWeb topic scorers and the Nemotron heads each form positively correlated blocks, and FineWeb-Edu aligns most closely with GneissWeb-Q, so the general-purpose "quality" classifiers largely share a common signal that DCLM does not; FineWeb2-HQ, the multilingual model, correlates weakly or negatively with the English-trained ones.

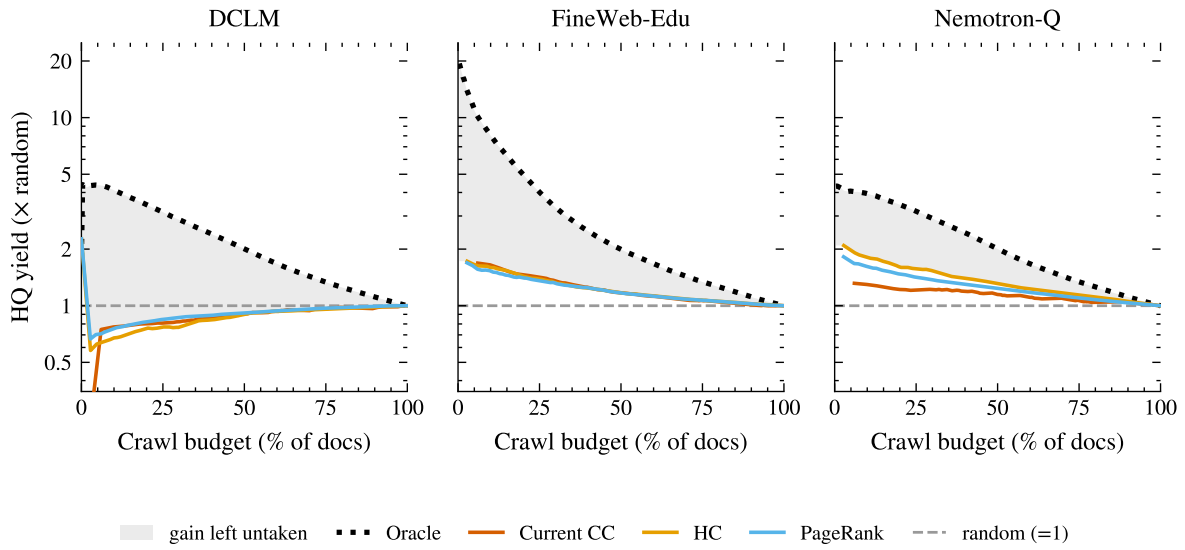


Figure 9: Full budget sweep behind Figure 2: high-quality documents collected as the crawl budget grows, divided by what a random crawl collects at the same budget (log-scaled y-axis; the dashed line at 1 means "no better than random"). Ordering hosts by harmonic centrality (HC), PageRank (PR), or Common Crawl's current allocation hugs the random line and stays far below the *oracle*; the shaded band shows the gap between centrality-based yield and the oracle upper bound, e.g, widest at small budgets, where crawl order matters most, and closing as the budget approaches a full crawl. For DCLM, centrality is even worse than random.

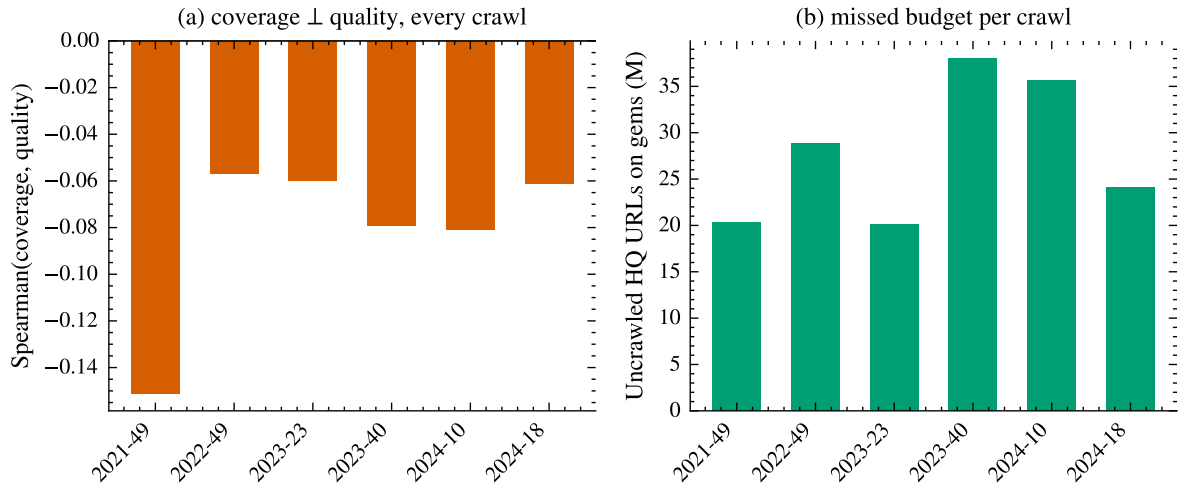


Figure 10: The coverage–quality disconnect across 6 Common Crawl snapshots (2021–2024). (a) The Spearman correlation between crawl coverage and host quality density is near zero and negative in every crawl (-0.057 to -0.151). (b) Each crawl leaves tens of millions of high-quality URLs uncrawled on hidden-gem hosts. The misallocation is a persistent, not a single-crawl, artifact—the plotted form of Table 1.

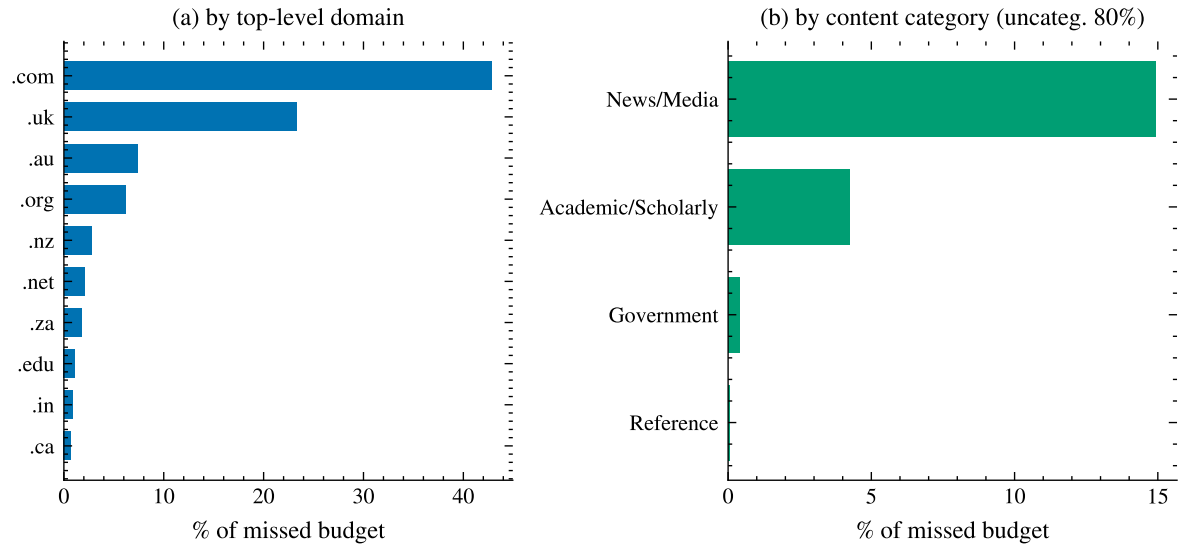


Figure 11: Where the missed high-quality budget sits, by (a) top-level domain and (b) content category (aggregates only; no host is named). Country-code publisher TLDs are over-represented relative to their host count—.uk alone holds 23% of the missed budget from just 6.1% of hosts at 34% median coverage. Among the hosts a keyword classifier can label, news/media and scholarly hosts dominate; the remaining ∼80% is uncategorised (a keyword lower bound), so panel (b) shows only the labelled share and omits the "Other" bucket. "Reference" denotes encyclopedic/reference sites (e.g. Wikipedia-family, dictionaries, encyclopedias); it is near zero because the major reference sites are already well crawled, so only a few dozen such hosts remain under-crawled gems.

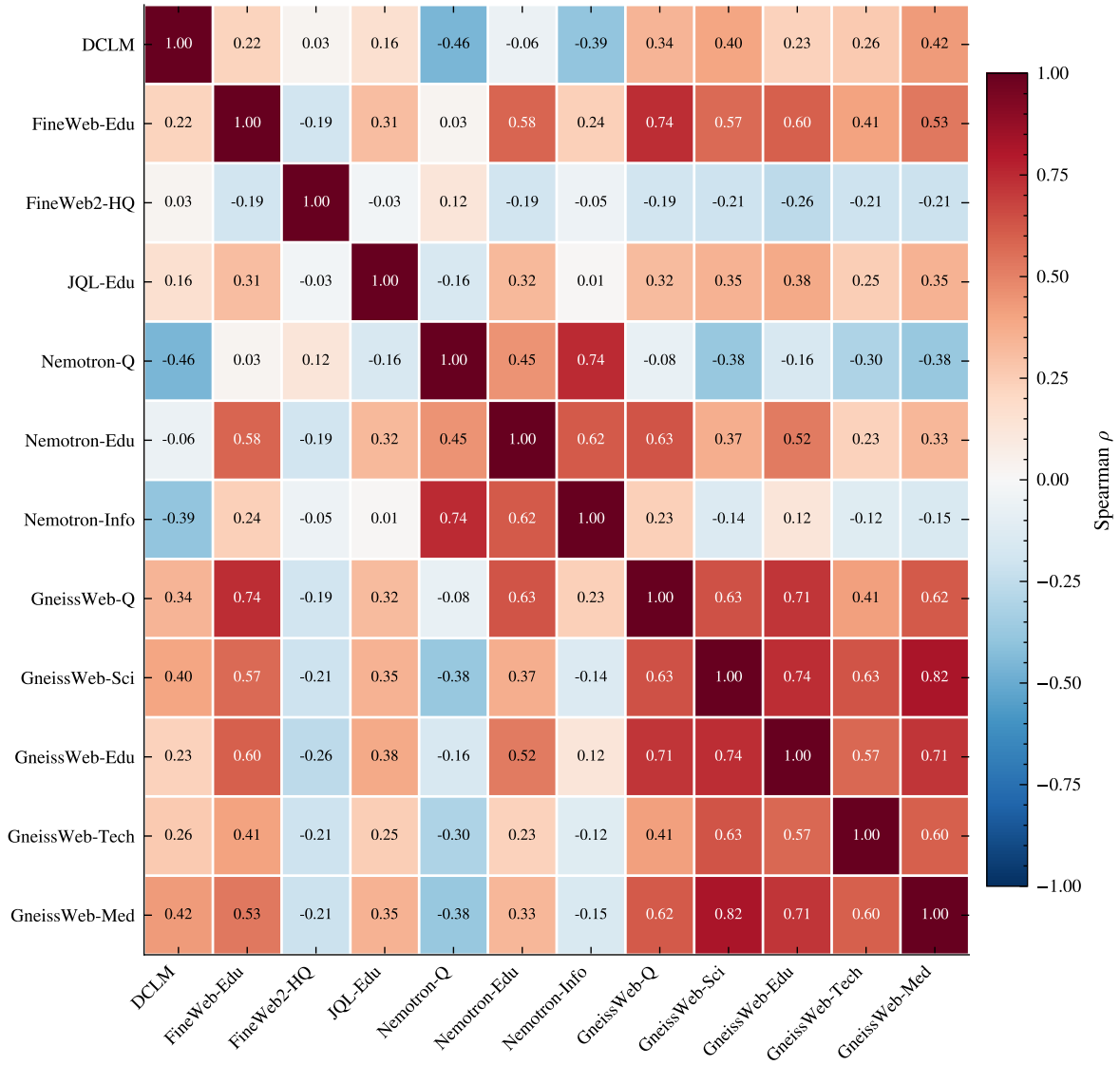


Figure 12: Pairwise Spearman rank correlation between the twelve classifiers' document scores (462,494 core documents; the diagonal is 1 by construction). The scale runs from -1 (blue) through 0 (white) to $+1$ (red). DCLM is largely orthogonal to the other quality signals and *negative* against the Nemotron heads, whereas the GneissWeb topic scorers and the Nemotron heads each form positively correlated blocks. This is the full pairwise view behind the single-classifier robustness of the coverage result (Figure 8).

Auditing Source Exposure in Baidu and Google AI Search

Yibo Li **Enci Guan** **Yuedan Cai** **Geng Liu** **Francesco Pierri**
Politecnico di Milano, Italy
{yibo2.li, enci.guan, yuedan.cai}@mail.polimi.it
{geng.liu, francesco.pierri}@polimi.it

Abstract

AI-generated overviews are becoming an increasingly prominent layer of search interfaces, yet their behavior in Chinese-language search remains underexplored. We conduct a cross-lingual audit of AI overview behavior on Baidu and Google using English queries sampled from MS MARCO and their translated Chinese counterparts. Our analysis examines when overviews are triggered across platform-language settings, which host domains receive visible exposure in Chinese-language overviews, how concentrated that exposure is, and how source overlap varies across settings. We also compare the embedding-based semantic similarity of generated answers for matched query intents. The results reveal substantial differences across platform-language settings in overview availability and visible source exposure. At the aggregate level, the settings exhibit low overlap in visible host-domain inventories, while matched-query answers yield median cosine similarities ranging from 0.701 to 0.813. These findings indicate that answer-level semantic similarity and aggregate source exposure capture distinct dimensions of AI-mediated search. Evaluations of AI search should therefore consider not only the content of generated answers but also how source visibility is distributed across platforms, languages, and information environments.

1 Introduction

Search engines have moved from document-ranking interfaces toward answer-producing interfaces (Kirsten et al., 2026; Grossman et al., 2026). Instead of returning only a ranked list of webpages, search result pages may now present an AI-generated overview above or alongside traditional search results. Such overviews typically summarize an answer and display a small set of visible source references (Liu et al., 2023). As a result, they function not only as summary components but also as a new layer of information selection (Huang et al., 2026). This shift raises an important measurement question. In traditional search, source visibility is largely determined by ranking position and the resulting user interaction (Joachims et al., 2005). AI overviews redistribute this visibility by displaying a selected set of sources alongside generated answers. Auditing AI search therefore requires examining not only the generated answer but also when overviews appear, which sources they reference, and how concentrated this source exposure is (Aral et al., 2026; Grossman et al., 2026). This matters because AI overviews mediate not only what information users read, but also which sources remain visible in the search interface. Differences in source exposure may therefore shape which websites and information providers users encounter, even when generated answers are semantically similar.

Existing empirical audits of generative search have focused primarily on Google and other English-dominant systems. Using user-side observations, these studies examine overview triggering, source selection, and citation exposure (Kirsten et al., 2026; Grossman et al., 2026; Hu et al., 2026; Aral et al., 2026). However, comparatively little is known about these patterns in Chinese-language search or how they vary across platforms and languages.

This gap is particularly relevant in cross-lingual settings, where query language may be associated with different web infrastructures and platform ecosystems. Queries expressing the same information need in different languages may therefore lead search systems to expose different sources, including domestic platforms, encyclopedic sources, institutional websites, social media, and commercial websites. Prior work has shown that source exposure in AI-mediated search can vary across systems, languages, and query formulations (Huang et al., 2026; Chen et al., 2025).

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 88–99
October 29, 2026 ©2026 Association for Computational Linguistics

We examine AI overviews as observable search-interface outputs, focusing on Chinese-language search. Rather than inferring the internal retrieval or generation mechanisms of Baidu and Google, we analyze whether an overview appears, how many source references it displays, which host domains receive visibility, and how source exposure varies across platform-language settings. Our primary comparison is between Baidu Chinese and Google Chinese, with Google English providing a cross-lingual reference. Baidu English is included only in the overall analysis of overview appearance because it produced too few successful cases for reliable downstream source analysis. Consequently, the study does not provide a fully crossed design for estimating independent platform and query-language effects. We therefore interpret the results as differences among the observed platform–language settings rather than as isolated effects of platform or language.

We address three research questions:

- **RQ1:** How frequently do AI overviews appear on Baidu and Google under Chinese- and English-query settings?
- **RQ2:** Which host domains are displayed in Chinese-language AI overviews on Baidu and Google, and how concentrated is their exposure?
- **RQ3:** How do visible host-domain overlap and answer-level semantic similarity vary across Baidu Chinese, Google Chinese, and Google English?

To address these questions, we collect 1998[1] query-platform observations from Baidu and Google. The base query set is drawn from MS MARCO (Nguyen et al., 2016), a large-scale search-oriented dataset constructed from real, anonymized Bing queries. MS MARCO provides realistic user information needs and a common set of search intents for aligned comparisons across platforms and languages. We retain the original English queries and translate them into Simplified Chinese, producing two approximately aligned language versions of each query intent. We analyze overview appearance across all four platform-language settings. Topic-level and source-oriented analyses focus on Baidu Chinese and Google Chinese, with Google English serving as a cross-lingual reference. We measure overview appearance, source volume, host-domain concentration, cross-setting host-domain overlap, and query-level semantic similarity between generated answers.

Overall, our findings show that different search settings can display semantically similar answers while displaying substantially different sets of visible host domains.

[1]The intended dataset contained 2000 observations, corresponding to 500 queries across four platform-language settings. One observation was not recorded in each Google setting, resulting in 499 collected records for Google Chinese and 499 for Google English. These missing observations were therefore not assigned to the Success, NoOverview, Blocked, or Error categories.

2 Related Work

Generative search and source exposure. Generative search systems synthesize retrieved information into answer-like outputs, creating a distinct layer of source selection and visibility. Studies of Google Search, Gemini, AI Overviews, and Featured Snippets show that generative search differs from traditional retrieval in its synthesis, stability, and presentation of sources (Kirsten et al., 2026; Grossman et al., 2026; Hu et al., 2026). AI-mediated search can also reshape which sources are exposed to users and affect downstream website traffic (Aral et al., 2026; Huang et al., 2026; Khosravi and Yoganarasimhan, 2026). Much of this literature focuses on Google and other Western or English-dominant systems, leaving Chinese-language search comparatively underexplored.

Citation support and source concentration. Prior work examines whether citations in generative search adequately support generated claims and how source exposure is distributed across websites (Liu et al., 2023; Yang, 2025). Other studies caution that visible citations do not necessarily guarantee reliable, high-quality, or non-synthetic evidence (Allaham and Diakopoulos, 2026; Narayanan Venkit et al., 2025). Our study does not assess claim-level support or source quality; instead, it treats visible host-domain frequencies as an interface-level measure of source exposure and concentration.

Cross-platform and cross-lingual search. Comparisons of Baidu and Google have identified platform- and language-dependent differences in autocompletion, moderation, and factual answers (Liu et al., 2024, 2025, 2026). Related work on generative engine optimization further examines how

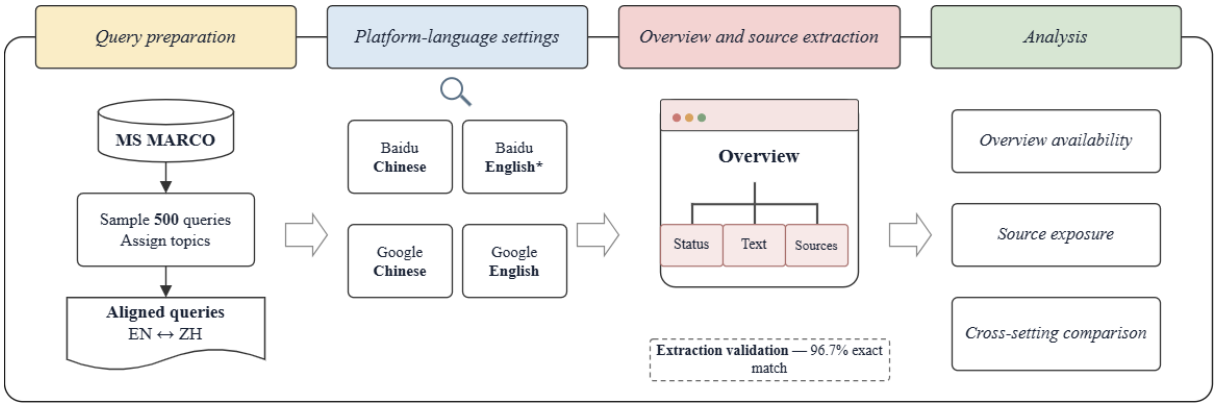


Figure 1: Overview of the audit pipeline. English queries are selected from MS MARCO and automatically translated into Simplified Chinese. The aligned queries are submitted to Baidu and Google, after which AI overviews, answer texts, and visible source domains are extracted for downstream analysis. Baidu English is included only in the overall overview appearance analysis.

content and source characteristics affect visibility in generated search responses (Chen et al., 2025; Aggarwal et al., 2024). Building on these studies, we conduct a user-side audit of overview triggering, host-domain concentration, source overlap, and answer-level similarity across Baidu Chinese, Google Chinese, and Google English.

3 Data and Methods

Figure 1 summarizes the data collection and analysis workflow, from query selection and translation to overview collection, source extraction, and metric computation.

3.1 Query Selection and Translation

The query set is drawn from MS MARCO v1.1, a search-oriented dataset containing real anonymized user queries originally submitted to Bing (Nguyen et al., 2016). We choose MS MARCO because its queries reflect realistic information needs and provide a common set of query intents that can be compared consistently across platforms and languages. After removing duplicates, we randomly sampled 500 queries without topic-based filtering to preserve the natural distribution of user information needs. We retain the original English queries and translate them into Simplified Chinese using Google Translate through the `deep_translator`[2] Python package. This produces aligned English and Chinese versions of each query. To assess translation quality, one author, a native Chinese speaker proficient in English, manually reviewed a random sample of 100 English–Chinese query pairs. A translation was judged acceptable when it preserved the primary information need of the original English query without substantially adding, omitting, or altering the requested information. All 100 sampled translations were judged acceptable. Before data collection, each English query was assigned exactly one of six topic labels using ChatGPT-5.5[3] and a fixed taxonomy adapted from Aral et al. (2026). The same label was assigned to the corresponding Chinese translation to preserve cross-lingual alignment. To assess annotation accuracy, one author manually reviewed 60 randomly sampled assignments, of which 96.7% were judged correct. The annotation prompt, category definitions, and illustrative examples are provided in Appendix E.

Both the original English queries and their Simplified Chinese translations were submitted to Baidu and Google, yielding four platform-language settings: Baidu Chinese, Baidu English, Google Chinese, and Google English. Data were collected between June 20 and July 10, 2026, using Google Chrome from Milan, Italy. The platform-language settings were collected in parallel over the same period, so matched query conditions were observed within approximately the same collection window. Google searches were performed while logged in to a Google account. The Chrome interface and Google Search language were set to English, whereas Baidu was accessed through its Chinese interface. No incognito mode, VPN, or proxy was used, and the same environment settings were kept fixed throughout the study. Search-result

[2] https://pypi.org/project/deep-translator/

[3] https://help.openai.com/en/articles/11909943-gpt-5-5-in-chatgpt

collection was automated using Playwright. Only source references visible in the rendered overview were recorded; no additional source panels were manually expanded. Queries were submitted sequentially, and the rendered search page, overview status, overview text, and visible source references were recorded for each query. Overall, we collected 1998 query-platform observations: 500 Baidu Chinese records, 500 Baidu English records, 499 Google Chinese records, and 499 Google English records. All four platform-language settings are included in the overall overview appearance analysis. Since Baidu English yielded only 20 successful AI overviews out of 500 queries, we exclude it from the topic-level and source-oriented analyses and report these cases descriptively in Appendix D.

Each record stores the query, topic label, collection status, and overview text when available. A record is labeled *Success* when the platform-specific extraction procedure detects a qualifying AI-overview component and extracts non-empty overview text. A qualifying overview was defined as a platform-specific AI-generated overview component detected in the rendered search-results page with non-empty generated answer text. The remaining records are classified as `NoOverview`, `Blocked`, or `Error`. `NoOverview` indicates that the results page was successfully rendered but no qualifying overview was detected, whereas `Blocked` and `Error` denote access restrictions and technical failures, respectively. Definitions, outcome counts, and representative examples are provided in Appendix C.

Figure 2 shows the topic distribution of the 500 base queries. The distribution is highly imbalanced: General Knowledge contains 325 queries, whereas Internet/Technology/Media contains only 10, and the remaining categories contain between 30 and 51 queries. Consequently, topic-level percentages for the smaller categories are based on relatively few observations and may be sensitive to individual records. We therefore report these results as descriptive breakdowns of the collected sample rather than as balanced comparisons across topics.

3.2 Source Extraction

For successful overview records, we extract visible source references and normalize them to lowercase host domains, such as `baike.baidu.com` or `en.wikipedia.org`. We group repeated references by host while treating distinct subdomains,

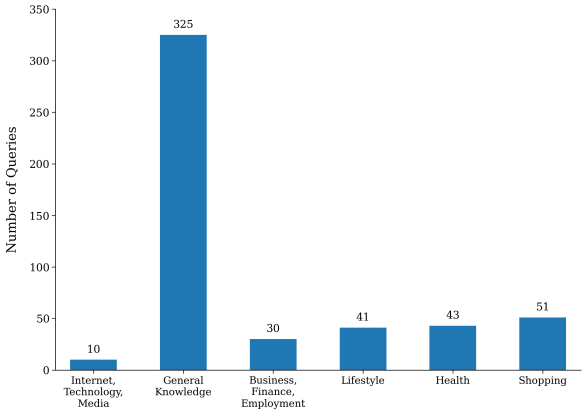


Figure 2: Topic distribution of the 500 base queries.

such as `baike.baidu.com`, `zhidao.baidu.com`, and `baijiahao.baidu.com` as separate hosts. Repeated appearances of the same host within an overview were deduplicated, so each host contributes at most one occurrence per overview. Accordingly, all reference counts reported below represent overview-level host occurrences rather than raw source-card or citation counts. We retain these subdomains separately because they correspond to distinct user-facing source labels and content services in the search interface. Accordingly, the results are interpreted at the host-domain level rather than at the level of registrable domains or corporate ownership.

We focus on host-domain exposure rather than exact URL exposure because the search interfaces do not always display full URLs consistently. Since Baidu and Google display source information differently, we use platform-specific extraction procedures. For Baidu, we extract displayed domains, primarily from the explicit source lists in the captured text. For Google, we map source labels and visible cards to host domains whenever possible. Therefore, cross-platform comparisons capture visible source exposure rather than providing a complete reconstruction of the search engines' internal evidence sets. To assess extraction accuracy, one author manually reviewed 60 randomly sampled successful overview records, including 20 from each retained source-analysis setting: Baidu Chinese, Google Chinese, and Google English. Extracted host domains exactly matched the visible source references in 58 of the 60 records, corresponding to an exact-match rate of 96.7%. Detailed validation results are provided in Appendix A.

3.3 Metrics

We define the overview appearance rate as

$$\mathrm{OAR} = \frac{\text{number of successful overview records}}{\text{number of collected records}}. \quad (1)$$

For each source-analysis setting, source volume is measured by the total number of overview-level host occurrences and the number of unique host domains. Let c_d denote the number of overview records in which host domain d appears. Its reference share is

$$s_d = \frac{c_d}{\sum_i c_i}, \quad (2)$$

where the denominator sums the reference counts over all host domains observed in the corresponding platform-language setting. We report the largest host-domain share (Top-1) and the Gini coefficient of the host-domain frequency distribution, following prior work on source concentration in AI search systems (Yang, 2025). Higher values indicate more concentrated source exposure. We quantify host-domain overlap between settings using Jaccard similarity:

$$J(A, B) = \frac{|D_A \cap D_B|}{|D_A \cup D_B|}, \quad (3)$$

where D_A and D_B denote the sets of unique host domains observed under settings A and B (Grossman et al., 2026).

As a complementary answer-level measure, we compute the cosine similarity between overview-text embeddings. To ensure comparability across the three pairwise analyses, we retain the common set of query intents for which all three settings generated an overview. The overview texts are encoded using `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2020). The encoder uses its default maximum sequence length of 128 tokens; longer overview texts are truncated during encoding.

For a matched query q and settings A and B, cosine similarity is defined as

$$\mathrm{Sim}_q(A, B) = \frac{\mathbf{e}_{A,q} \cdot \mathbf{e}_{B,q}}{\|\mathbf{e}_{A,q}\| \|\mathbf{e}_{B,q}\|}, \quad (4)$$

where $\mathbf{e}_{A,q}$ and $\mathbf{e}_{B,q}$ denote the corresponding overview-text embeddings. Higher values indicate greater semantic similarity but do not imply factual correctness, citation support, or source overlap. Because host-domain Jaccard is computed over setting-level aggregate host sets, whereas semantic similarity is computed over matched query pairs, we treat the two metrics as complementary descriptive measures. We therefore do not interpret their coexistence as evidence of a query-level relationship between answer similarity and source overlap.

4 Results

4.1 Overview Appearance Rate

Table 1 reports the overall overview appearance rates across the four platform-language settings. The *Other* category combines `NoOverview`, `Blocked`, and `Error` records; their definitions and disaggregated counts are provided in Appendix C.

Google generated overviews for 97.6% of Chinese queries and 97.0% of English queries. The corresponding rates for Baidu were 82.0% and 4.0%, respectively. Overview appearance therefore varied substantially across the observed platform–language settings, with the largest contrast occurring between the Baidu Chinese and Baidu English settings. As a sensitivity check, excluding Blocked and Error records from the denominator yielded corresponding OARs of 82.0%, 4.0%, 97.8%, and 97.6%, respectively, leaving the overall pattern unchanged.

Setting	Total	Success	Other	OAR
Baidu Chinese	500	410	90	82.0%
Baidu English	500	20	480	4.0%
Google Chinese	499	487	12	97.6%
Google English	499	484	15	97.0%

Table 1: Overview appearance rate across platform-language settings.

Baidu exhibited a pronounced language difference, generating overviews for 82.0% of Chinese queries but only 4.0% of English queries. Because Baidu English produced only 20 successful overviews, we retain it in the overall appearance analysis but exclude it from the topic-level and source-oriented analyses. Figure 3 therefore reports topic-level results for Baidu Chinese, Google Chinese, and Google English.

At the topic level, Google maintained high overview appearance rates across all categories, ranging from 90.2% to 100.0% in both language settings. Baidu Chinese exhibited greater variation, with the highest overview appearance rate for Health (97.7%) and the lowest rate for Shopping (70.6%). Rates for the remaining topics ranged from 80.0% to 83.3%.

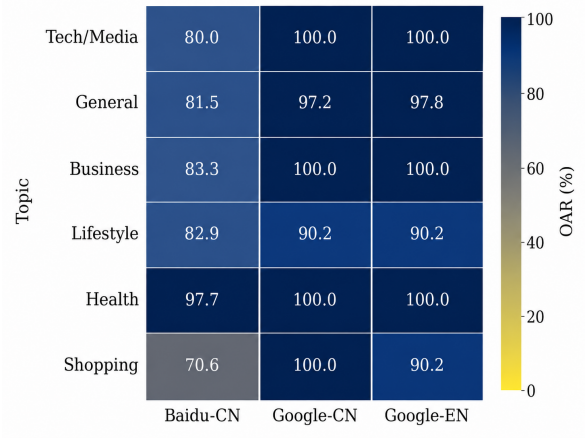


Figure 3: Topic-level overview appearance rates for Baidu Chinese, Google Chinese, and Google English. Each cell reports the percentage of queries within a topic that produced an AI overview.

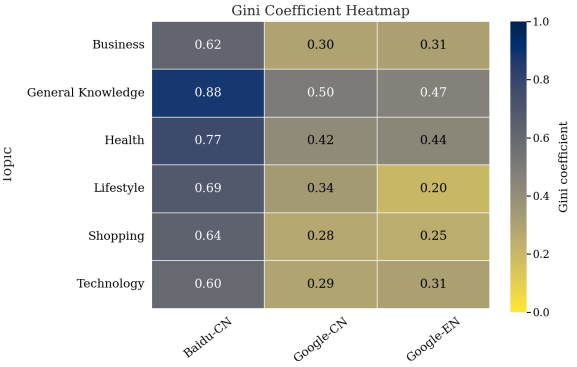


Figure 4: Topic-dependent source concentration measured by host-domain Gini coefficients. Baidu Chinese remains more concentrated than Google Chinese and Google English across most topics.

Pair	Shared	Union	Jaccard
G-CN / G-EN	179	5,341	0.034
B-CN / G-CN	103	3,183	0.032
B-CN / G-EN	25	3,095	0.008

Table 2: Host-domain overlap between source-analysis settings. B-CN, G-CN, and G-EN denote Baidu Chinese, Google Chinese, and Google English.

4.2 Source Volume and Concentration

Detailed source-volume and host-domain concentration statistics across the three source-analysis settings are reported in Appendix B, Table 3, while Figure 4 reports the corresponding topic-level Gini coefficients.

Baidu Chinese contained fewer overview-level host occurrences and unique host domains than the two Google settings. After normalizing by the number of successful overview records, the mean number of distinct host occurrences per overview was 15.0 for Baidu Chinese, 18.6 for Google Chinese, and 16.7 for Google English. Thus, the lower aggregate host-occurrence count for Baidu Chinese was also reflected at the per-overview level, although the difference was smaller after normalization. However, its source exposure was substantially more concentrated: its Gini coefficient was 0.895, compared with 0.500 for Google Chinese and 0.462 for Google English. Its Top-1 host-domain share was also higher (27.1%, compared with 6.6% and 5.5%, respectively).

The topic-level results show a similar pattern, with Baidu Chinese exhibiting higher host-domain concentration than Google Chinese and Google English across most topics.

4.3 Top Referenced Host Domains

Figure 5 shows the most frequently extracted host domains in each source-analysis setting. For Baidu Chinese, the most frequently referenced host domains include `baike.baidu.com`, `baijiahao.baidu.com`, and `zhidao.baidu.com`. These are distinct host domains under baidu.com, while the overall host-domain frequency distribution is substantially more concentrated than in the two Google settings.

Google Chinese and Google English exhibit a more diverse source composition. Their most frequently referenced domains include `zh.wikipedia.org` and `mayoclinic.org` in Chinese, and `en.wikipedia.org` and `sciencedirect.com` in English. Overall, these frequency distributions are consistent with the concentration results reported above: visible source exposure is more concentrated for Baidu Chinese and more dispersed across external hosts for both Google settings.

4.4 Cross-Platform Domain Overlap

Table 2 reports Jaccard similarity between extracted host-domain sets. Overlap is low across all three comparisons. Google Chinese and Google English have the highest similarity (0.034), followed by Baidu Chinese and Google Chinese (0.032). Baidu Chinese and Google English have the lowest similarity (0.008). Thus, the aggregate host-domain inventories exposed by the three settings show little overlap.

Jaccard similarity captures set overlap but does not account for domain frequency. For example, `baike.baidu.com` appears in both Baidu Chinese

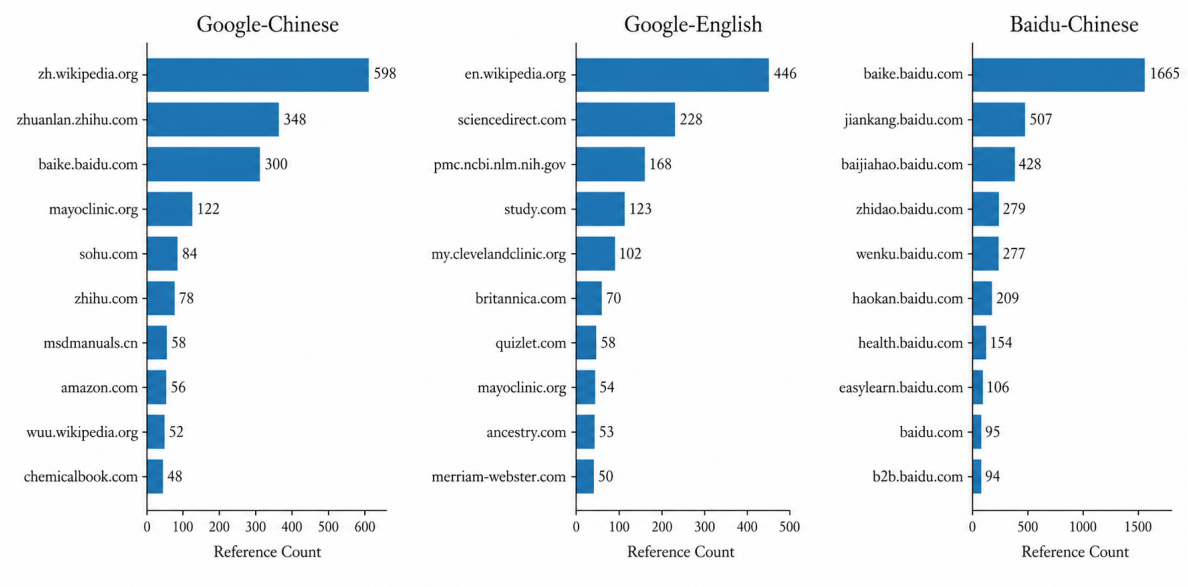


Figure 5: Overall top-referenced host domains across source-analysis settings. Baidu Chinese is dominated by Baidu-owned host domains, while Google Chinese and Google English expose a more mixed set of sources. Note that the x-axis scales differ across panels because reference counts differ substantially across settings.

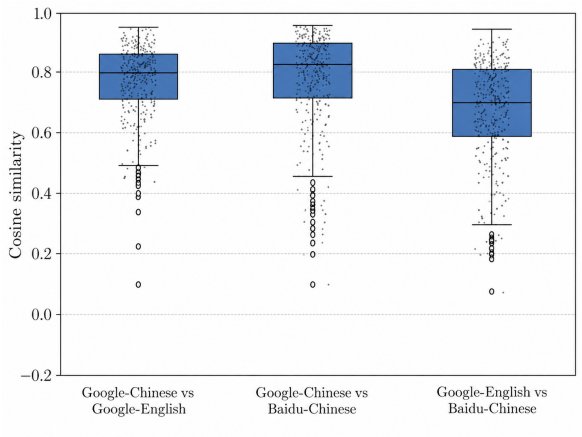


Figure 6: Query-level cosine similarity between generated answers for matched query intents. Only the common set of query intents for which all three settings successfully generated an overview is included. G-CN, G-EN, and B-CN denote Google Chinese, Google English, and Baidu Chinese, respectively. Boxes show the interquartile range, center lines indicate medians, and points represent matched queries.

and Google Chinese, although the two settings share only 103 of their 3183 combined host domains. Thus, individual domains may be shared even when the overall overlap between the two domain sets is limited. These Jaccard values should be interpreted as descriptive measures of aggregate host-set overlap. Because the metric is unweighted and sensitive to the number of low-frequency hosts in each setting, low values do not by themselves establish unusually low overlap relative to a statistical null model.

4.5 Semantic Similarity of Generated Answers

Figure 6 reports the query-level cosine similarity between generated answers for three pairwise comparisons: Google Chinese versus Google English, Google Chinese versus Baidu Chinese, and Google English versus Baidu Chinese. To ensure comparability across the three pairwise analyses, we restrict the analysis to the common set of 331 query intents for which all three settings successfully generated an overview.

The three comparisons yield median cosine similarities ranging from 0.701 to 0.813, although their distributions differ. Google Chinese and Baidu Chinese show the highest median cosine similarity at 0.813 (IQR: 0.709–0.886), followed by Google Chinese and Google English at 0.791 (IQR: 0.707–0.850). Google English and Baidu Chinese show a lower median similarity of 0.701 (IQR: 0.589–0.803) and a wider distribution, suggesting greater answer-level variation when both the platform and query language differ.

At the aggregate level, these answer-similarity results coexist with the low host-domain Jaccard similarities reported in Section 4.4. Because the two metrics operate at different levels of aggregation, we do not interpret this pattern as evidence of

a query-level relationship between semantic similarity and source overlap. Instead, the results show that answer similarity and aggregate source exposure capture distinct dimensions of AI-mediated search.

5 Discussion and Conclusion

AI-generated overviews shift search interfaces from ranked lists of webpages toward answer-first presentations that also determine which sources remain visible to users. Our results show that this shift is not uniform across the observed platform–language settings. Overview availability, source concentration, and visible source overlap vary substantially across settings, even when the queries represent the same underlying information needs. The central takeaway is therefore methodological: answer content and visible source exposure provide complementary views of AI-mediated search and should not be treated as interchangeable indicators.

RQ1: ***How frequently do AI overviews appear on Baidu and Google under Chinese- and English-query settings?*** Google generated overviews for nearly all collected queries, with appearance rates of 97.6% for Chinese queries and 97.0% for English queries. Baidu showed a substantially stronger language difference: its appearance rate was 82.0% for Chinese queries but only 4.0% for English queries. These results show that overview availability varied substantially across the observed platform–language settings. The very small number of successful Baidu English cases also means that this setting cannot support reliable topic-level or source-oriented comparisons.

RQ2: ***Which host domains are displayed in Chinese-language AI overviews on Baidu and Google, and how concentrated is their exposure?*** Baidu Chinese displayed fewer overview-level host occurrences and unique host domains than Google Chinese, while its source exposure was considerably more concentrated. Its host-domain-level Gini coefficient was 0.895, compared with 0.500 for Google Chinese, and its most frequently referenced host accounted for 27.1% of all host occurrences, compared with 6.6% for Google Chinese. The most visible Baidu Chinese sources were also primarily Baidu-owned subdomains. In contrast, Google Chinese exposed a broader mixture of encyclopedic, medical, commercial, and other external host domains. These findings show that displaying multiple source references does not necessarily imply broad source diversity.

RQ3: ***How do visible host-domain overlap and answer-level semantic similarity vary across Baidu Chinese, Google Chinese, and Google English?***

Host-domain overlap was low across all three pairwise comparisons. The largest Jaccard similarity was 0.034 between Google Chinese and Google English, whereas Baidu Chinese and Google English had a similarity of 0.008. In contrast, matched-query answer comparisons yielded median cosine similarities ranging from 0.701 to 0.813. Because these measures operate at different levels of aggregation, the results should not be interpreted as evidence of a query-level relationship between answer similarity and source overlap. Instead, they show that answer-level similarity and aggregate source exposure capture distinct dimensions of AI-mediated search.

More broadly, these findings show that answer-level similarity alone does not provide a complete account of AI-mediated search. Across the observed settings, median matched-query cosine similarity ranged from 0.701 to 0.813, while aggregate host-domain inventories showed low overlap. Audits of AI search should therefore consider not only whether an overview appears and what answer it contains, but also which sources receive visibility, how concentrated that exposure is, and how these patterns vary across platforms and languages.

This distinction is particularly important for Chinese-language search, where Baidu Chinese showed a more concentrated host-domain distribution than Google Chinese, while several of its most frequently referenced hosts were Baidu subdomains. Future work could extend this audit using independently collected native-language queries, repeated observations across locations and time periods, and evaluations of source relevance, credibility, claim-level support, and downstream user behavior.

6 Limitations

This study is limited by the size and timing of the collected sample. Search results may change over time, and different locations or browser states may produce different pages. The extraction process also depends on page structure. If a search engine changes how it displays overview sources, the extraction rules may need to be updated. In addition, Google searches were conducted while logged in to a Google account and with the interface language set to English, so account state and interface language may have influenced the observed results. In particular, collecting Baidu from Milan may not reproduce the search experience of users accessing the service from mainland China. The observed Baidu English overview rate may reflect regional routing, interface behavior, language support, or a combination of these factors.

Another limitation comes from query translation. The Chinese query set is produced by automatically translating the original English MS MARCO queries. Although this keeps the underlying query intents approximately aligned across languages, translation may change wording, ambiguity, or cultural context. Therefore, the cross-lingual comparison should be interpreted as a comparison between original English queries and their translated Chinese counterparts, rather than between two independently collected native query sets. The topic distribution is also highly imbalanced, so topic-level results are intended as descriptive breakdowns and should not be interpreted as equally precise comparisons across categories.

The host-domain-level analysis does not judge whether a source is correct, relevant, or trustworthy. A host domain can be frequently cited while still providing weak support for a specific overview statement. Conversely, a low-frequency source may be highly relevant for a particular query. Therefore, reference frequency should be interpreted as a visibility measure rather than a quality measure. Because the analysis preserves user-facing subdomains as distinct hosts, organization-level aggregation could yield different concentration and overlap estimates, particularly for platforms that expose multiple services under the same corporate domain. The reported Gini and Jaccard values should therefore be interpreted as descriptive statistics under our host-domain definition. Alternative aggregation units, such as registrable domains or corporate ownership, as well as alternative overlap or concentration measures, could produce different estimates. Future work should examine the robustness of these patterns under such alternative specifications.

Ethics Statement

This study analyzes publicly visible search-engine interface outputs and does not involve human participants or personally identifiable information. The queries are drawn from the anonymized MS MARCO dataset and are used only to compare platform behavior. All findings are reported in aggregate.

The study does not attempt to identify individual users, infer private system internals, or determine the cultural composition of model training data. It does not classify websites as trustworthy or untrustworthy, since source frequency measures interface-level visibility rather than source quality. The findings describe the observed collection period rather than a permanent assessment of either platform. The collection process did not attempt to bypass access restrictions or verification mechanisms; such cases were recorded as Blocked. Source-exposure measurements may also be useful for generative engine optimization. We therefore frame the results as an audit of aggregate visibility patterns rather than as actionable guidance for manipulating source rankings or platform exposure.

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5–16, New York, NY, USA. Association for Computing Machinery.

Mowafak Allaham and Nicholas Diakopoulos. 2026. Synthetic sources?: Auditing generative search engine citations for evidence of ai-generated sources. *Preprint*, arXiv:2605.23684.

Sinan Aral, Haiwen Li, and Rui Zuo. 2026. The rise of ai search: Implications for information markets and human judgement at scale. *Preprint*, arXiv:2602.13415.

Mahe Chen, Xiaoxuan Wang, Kaiwen Chen, and Nick Koudas. 2025. Generative engine optimization: How to dominate ai search. *Preprint*, arXiv:2509.08919.

Riley Grossman, Songjiang Liu, Michael K. Chen, Mike Smith, Cristian Borcea, and Yi Chen. 2026. How generative ai disrupts search: An empirical study of

google search, gemini, and ai overviews. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, page 448–459, New York, NY, USA. Association for Computing Machinery.

Desheng Hu, Joachim Baumann, Aleksandra Urman, Elsa Lichtenegger, Robin Forsberg, Anikó Hannák, and Christo Wilson. 2026. Auditing google's ai overviews and featured snippets: A case study on baby care and pregnancy. *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1):1044–1062.

Michelle Huang, Agam Goyal, Koustuv Saha, and Eshwar Chandrasekharan. 2026. Answer bubbles: Information exposure in ai-mediated search. *Preprint*, arXiv:2603.16138.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 154–161, New York, NY, USA. Association for Computing Machinery.

Mehrzad Khosravi and Hema Yoganarasimhan. 2026. Impact of ai search summaries on website traffic: Evidence from google ai overviews and wikipedia. *Preprint*, arXiv:2602.18455.

Elisabeth Kirsten, Jost Große Perdekamp, Qinyuan Wu, Mihir Upadhyay, Krishna P. Gummadi, and Muhammad Bilal Zafar. 2026. Characterizing web search in the age of generative AI. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 10827–10848, San Diego, California, United States. Association for Computational Linguistics.

Geng Liu, Carlo Alberto Bono, and Francesco Pierri. 2025. Comparing diversity, negativity, and stereotypes in chinese-language ai technologies: an investigation of baidu, ernie and qwen. *PeerJ Computer Science*, 11:e2694.

Geng Liu, Li Feng, Mengxiao Zhu, and Francesco Pierri. 2026. Evaluating reliability asymmetries in chinese factual search and ai answers. *Preprint*, arXiv:2602.22221.

Geng Liu, Pietro Pinoli, Stefano Ceri, and Francesco Pierri. 2024. A comparison of online search engine autocompletion in google and baidu. In *Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. Search engines in the ai era: A qualitative understanding to the false promise of factual and verifiable source-cited responses in llm-based search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1325–1340, New York, NY, USA. Association for Computing Machinery.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Kai-Cheng Yang. 2025. News source citing patterns in ai search systems. *Preprint*, arXiv:2507.05301.

Setting	Refs.	Hosts	Top-1	Gini
Baidu Chinese	6,154	443	27.1%	0.895
Google Chinese	9,053	2,843	6.6%	0.500
Google English	8,061	2,677	5.5%	0.462

Table 3: Source volume and concentration at the host-domain level. Baidu English is excluded because it produced only 20 successful overview records.

Setting	Success	NoOvr.	Blocked	Error	Total
Baidu Chinese	410	90	0	0	500
Baidu English	20	479	0	1	500
Google Chinese	487	11	1	0	499
Google English	484	12	3	0	499

Table 4: Distribution of successful and non-successful outcomes across platform-language settings. *NoOvr.* denotes *NoOverview*.

A Source-Extraction Validation

We randomly sampled 60 successful overview records for manual validation, with 20 records drawn from each retained source-analysis setting: Baidu Chinese, Google Chinese, and Google English. For each record, the extracted host-domain list was compared with the source references visibly displayed in the captured overview. A record was counted as an exact match only when all visible host domains were correctly extracted and no unsupported domain was included.

Overall, 58 of the 60 sampled records were exact matches, yielding an exact-match rate of 96.7%. The remaining two records contained extraction discrepancies. The validation was conducted by one author; therefore, inter-annotator agreement is not applicable.

B Detailed Source Volume and Concentration Statistics

Table 3 reports the total number of extracted source references, the number of unique host domains, the share of the most frequently referenced host domain, and the host-domain-level Gini coefficient for each retained source-analysis setting.

C Classification of Non-Successful Records

Records that were not labeled *Success* were assigned to three mutually exclusive categories: *NoOverview*, *Blocked*, and *Error*. The distribution of these outcomes across the four platform-language settings is reported in Table 4.

NoOverview indicates that the search-results page was successfully retrieved and rendered, but no qualifying AI-overview component was detected. This category therefore represents an observed absence of an overview rather than a data-collection failure.

Blocked indicates that access to the search-results page was restricted, preventing reliable observation of whether an AI overview was present.

Error indicates that the record could not be reliably classified because of a technical failure during data collection. The only observed error occurred when page loading exceeded the configured 25-second timeout.

Thus, *NoOverview* represents an observed absence of an AI overview, whereas *Blocked* and *Error* represent cases in which overview presence could not be reliably determined.

Representative page-level examples illustrate the distinction among these categories. A *NoOverview* record contains a normally rendered search-results page with standard results but no detected AI-overview component. A *Blocked* record instead displays an access-restriction or verification page rather than usable search results. The single *Error* record corresponds to a page that did not complete loading within the 25-second timeout.

D Descriptive Results for Baidu English

As shown in Table 5, Baidu English produced 20 successful AI overviews among the 500 submitted queries. Because of this small sample, the results below are reported descriptively and are not used for cross-setting comparisons.

Topic	Successful cases
Internet/Technology/Media	2
General Knowledge	9
Business/Finance/Employment	1
Lifestyle	1
Health	1
Shopping	6
Total	20

Table 5: Topic distribution of successful Baidu English overview cases.

E Topic Annotation Procedure

All topic labels were generated automatically using GPT-5.5 through the ChatGPT interface. Human involvement was limited to validating a random sample of the automatically assigned labels and correcting the two mismatched labels identified

Category	Definition	Example Queries

Internet/Technology/Media	Software, hardware, digital devices, online services, internet platforms, technology, news, and media.	`how to transfer photos and videos from android to computer`
		`what is a web domain`
		`who invented the first mechanical computer`

General Knowledge	General factual or informational queries that are not more appropriately covered by another domain-specific category.	`what are electromagnetic waves made of`
		`where is the amur leopard found`
		`what is a apothem`

Business/Finance /Employment	Companies, markets, banking, personal finance, careers, occupations, and employment.	`what qualifications do a police officer need`
		`average salary of a teacher in texas`
		`when must ira distributions begin`

Lifestyle	Travel, food, home, leisure, hobbies, relationships, and everyday activities.	`closest airport to larissa greece`
		`how long to steam artichokes in steamer`
		`weather in oregon in may`

Health	Diseases, symptoms, diagnosis, treatment, medicine, healthcare, and wellbeing.	`what causes inflamed pancreas`
		`how often should men get a PSA test`
		`what types of bacteria are found in the mouth`

Shopping	Products, brands, prices, purchasing decisions, availability, and product comparisons.	`is it better to buy a car from a dealer or owner`
		`average cost for custom canopy`
		`cost to replace brakes`

Table 6: Topic-category definitions and example queries. All examples are drawn from the English query dataset used for topic annotation and are reproduced verbatim, including any grammatical errors in the original queries.

during validation. Each original English query was assigned exactly one topic label. The same label was then assigned to its Simplified Chinese translation, rather than annotating the translated query independently, to preserve cross-lingual alignment.

E.1 Annotation Prompt

The same prompt and category taxonomy were used for all queries. The model was instructed to identify the primary information need and return exactly one of the six permitted category labels.

The following prompt template was used:

> Assign the following search query to exactly one of the categories below according to its primary information need.
>
> - Internet/Technology/Media
> - General Knowledge
> - Business/Finance/Employment
> - Lifestyle
> - Health
> - Shopping
>
> Select the single best-fitting category. If more than one category appears relevant, choose the category that most directly represents the user's main information need. Return only the category label and do not provide an explanation.
>
> **Search query:** `{query}`

In each classification request, `{query}` was replaced by the actual English query being annotated.

E.2 Category Definitions and Examples

The six categories were operationalized as shown in Table 6. The examples are drawn from the query dataset and illustrate the primary information need associated with each category.

E.3 Manual Validation

To assess the accuracy of the automatic topic annotation, one author manually reviewed a random sample of 60 labeled English queries. For each sampled query, the reviewer assessed whether the assigned category correctly represented its primary information need according to the definitions above.

Among the 60 reviewed assignments, 58 were judged correct, corresponding to an accuracy of 96.7%. The remaining two assignments were considered mismatched and were corrected before downstream analysis. Because the validation was conducted by one author, inter-annotator agreement was not computed.

Lost in Translation: Structured-Span Corruption in Machine-Translated Multilingual Web Corpora

Ruiyang Zhang
FLock.io & Ryonix Labs Inc.
zhangruiyang36@gmail.com

Abstract

Machine translation of English web text is a default recipe for building multilingual and low-resource corpora. We show that this recipe systematically damages the *structured spans* embedded in web text—URLs, email addresses, numbers and identifiers, code, markup and quotation delimiters—and that fluency metrics do not reveal the damage. Translating 56,163 span-bearing segments from a CC-MAIN-2026-30 Common Crawl sample into nine typologically diverse languages with NLLB-200, we find that 12.6% of 73,881 structured spans do not survive, ranging from 91.1% for numbers to 50.6% for quotation delimiters. Damage is strongly language-dependent (18.7 points), but in the *opposite* direction to the obvious prediction: the highest-resource targets are damaged most, an inversion that survives controlling for the systems' tendency to copy low-resource input through untranslated. Round-trip fluency barely tracks any of this ($\rho = 0.48$ across languages, $p = 0.19$; $\rho = 0.25$ per segment). We pre-registered the study, report the refuted hypothesis alongside the supported ones, and release the annotated parallel-span corpus with an independent re-derivation of every number.

1 Introduction

For most of the world's languages there is not enough native web text to build a large corpus, so a common substitute is to machine-translate English web text into the target language. The resulting corpora are then filtered and validated with corpus-level quality metrics: BLEU (Papineni et al., 2002), chrF (Popović, 2015), or learned estimators such as COMET (Rei et al., 2020) and its reference-free variant (Rei et al., 2022).

Those metrics score a translation by comparing surface forms in bulk—against a reference, or, when none exists, against the source via back-translation. Web text, however, is not only language. It is dense with *structured spans* that carry meaning through their exact surface form: a URL, an email address, a product code, a version string, an HTML entity, a quoted phrase. A translation can score well and still have silently rewritten a URL, converted digits to another numeral system, or dropped a quotation mark. Aggregate surface metrics are, by construction, nearly blind to this: a handful of altered characters barely moves chrF, but it is the difference between a resolvable URL and a dead one.

We quantify that gap. From a CC-MAIN-2026-30 Common Crawl WET sample we draw 5,000 English documents, extract 73,881 structured spans of six frozen types with documented detectors, and translate the 56,163 span-bearing segments into nine languages with NLLB-200 (NLLB Team et al., 2024), plus a second, LLM-based system as a robustness check.

Three hypotheses were registered before any translation was run. **H1** (span damage): translation does not preserve embedded structured spans, and the loss rate depends on the span type. **H2** (language dependence): preservation varies across target languages, with lower-resourced targets losing more. **H3** (blindness): corpus-level quality scores do not predict span damage. Each was given a falsification criterion in advance. H1 and H3 are supported; **H2 is refuted, and inverted**—the highest-resource targets are damaged most. We report the refutation rather than reframe it (§4).

Our contributions are:

- A **span-typed, language-dependent corruption measurement** on real web text: preservation rates by span type and target language.
- A **corruption taxonomy** that names how spans break—deletion, script transliteration, digit alteration, truncation, casing, delimiter substitution—and its distribution across languages.
- The **blindness result**: the relationship (or lack

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 100–106
October 29, 2026 ©2026 Association for Computational Linguistics

of one) between round-trip chrF2—source against back-translation—and span preservation, reported as measured.

- A **released annotated parallel-span resource**—source span, translated segment, preservation verdict and corruption label—together with the frozen detectors, the scoring code, and an independent verifier that rederives every number in this paper from the raw translations.

The study was pre-registered before any translation was run; the pre-registration, a deviations log, and all code are released with the resource.[1]

2 Related Work

Web corpora and their quality. Building corpora from the web is the founding concern of this workshop's community (Kilgarriff and Grefenstette, 2003; Baroni et al., 2009). Common Crawl now underpins C4 (Raffel et al., 2020), CCNet (Wenzek et al., 2020), OSCAR (Ortiz Suárez et al., 2020) and FineWeb (Penedo et al., 2024). Quality auditing of the multilingual portions of these resources has found serious problems: Kreutzer et al. (2022) audit web-crawled multilingual datasets and report substantial fractions of mislabelled or non-linguistic content, and MADLAD-400 (Kudugunta et al., 2023) documents extensive manual auditing. That line of work asks whether the *text* is good; we ask whether the *non-linguistic material inside* otherwise good text survives.

MT quality and its failure modes. Low-resource MT quality is uneven even in strong multilingual systems (NLLB Team et al., 2024), and hallucination and omission are well documented (Guerreiro et al., 2023). Round-trip translation (Sennrich et al., 2016) is a common cheap proxy for adequacy. Our concern is orthogonal to overall adequacy: a segment can be adequately translated as language while its embedded spans are destroyed.

Data documentation. Data statements (Bender and Friedman, 2018) and datasheets (Gebru et al., 2021) argue that resources should ship with explicit accounts of provenance and limitation. We propose span preservation as a concrete, cheap statistic that MT-built corpora should report alongside fluency.

[1]Corpus, detectors, scoring code, pre-registration, deviations log and the independent verifier: `https://github.com/ryonzhang/wac13-span-corruption`

3 Method

Corpus. We sample the CC-MAIN-2026-30 Common Crawl **WET** release (extracted plain text), using 100 WET files chosen by seeded random selection from the snapshot manifest, and 5,000 documents drawn by seeded reservoir sampling within those files. We keep records whose first `WARC-Identified-Content-Language` is `eng`, of 500–5000 characters, containing at least one detected span, deduplicated on normalised content hash.

We use WET rather than WARC deliberately. WET is what MT-based corpus pipelines actually consume, it is reproducible from a snapshot id plus a seed with no HTML-extraction confound, and code and markup still occur in it as literal text in tutorials, documentation and forum posts. The cost is that WET discards page structure, so "markup" here means markup-as-content rather than page scaffolding (§5).

Span detectors. We detect six types—URL, EMAIL, NUMBER, CODE, MARKUP, QUOTE—with deterministic regular expressions plus validators. Spans are non-overlapping, resolved by a fixed priority order, so `<a href="http://x.com">` is a single MARKUP span rather than a MARKUP/URL/QUOTE pile-up. CODE covers source-code fragments and code-like identifiers (snake_case, dotted paths, call forms); NUMBER requires at least three characters so that trivially preserved single digits do not dominate.

Detectors were developed on a 400-document development split drawn from WET files disjoint from the evaluation files. We manually adjudicated a stratified sample of 204 detected spans and 25 exhaustively hand-marked windows, made a single revision pass, then froze the detectors. Measured precision is 98.5% and recall 91.4% overall (Table 2).

Unit of analysis. The unit is the *span-bearing segment*. Documents are split at newlines and sentence ends—never inside a detected span—and hard-capped at 350 characters at a span-safe boundary. Because NLLB is a sentence-level system, segments are translated independently of their neighbours, so restricting translation to span-bearing segments produces output for those segments that is identical to translating the whole corpus, at roughly one eighth of the compute. Since our splitter never

cuts a span, any breakage we observe is attributable to translation rather than to preprocessing.

Systems. The primary system is NLLB-200-distilled-600M with deterministic decoding (beam 4, no sampling), translating all 43,981 unique segments into nine languages: French, German, Chinese, Arabic, Hindi, Swahili, Icelandic, Maltese and Yoruba. As a robustness check against system-specific behaviour we additionally translate a 10% seeded subsample with an instruction-tuned LLM (Llama-3.3-70B-Instruct, temperature 0).

Matching. Both sides are NFKC-normalised, stripped of zero-width characters and whitespace-collapsed; URL and EMAIL are compared case-insensitively. Every normalisation choice is deliberately lenient—it can only make a span easier to count as preserved—so the damage we report is a lower bound. For the five verbatim types, a span occurrence is preserved if and only if its normalised form occurs in the normalised translation, counted as $\min(\text{count}_{\text{src}}, \text{count}_{\text{tgt}})$.

QUOTE is scored differently, on **delimiter integrity**: the text inside a quotation is natural language and is supposed to change, so requiring verbatim preservation would be wrong. A QUOTE span counts as preserved if and only if a matched delimiter pair of the same class survives.

Metrics. We report (i) preservation by span type $\times$ language with Wilson intervals, plus a script-normalised variant that folds non-ASCII digits; (ii) a corruption taxonomy over non-preserved spans, assigned by deterministic rules in fixed priority order; (iii) fluency as chrF2 between the source segment and its NLLB round-trip back-translation, restricted to segments of at least eight words, together with a span-masked variant; and (iv) a parse proxy measuring whether a span is still recoverable and syntactically valid after translation.

4 Results

Span damage is large and strongly type-dependent. Across 73,881 spans and nine languages, 12.6% of span occurrences do not survive translation. The rate depends far more on span type than any fluency score would suggest: NUMBER survives at 91.1% and EMAIL at 84.1%, but CODE at only 60.0% and QUOTE at 50.6%. H1 is supported.

QUOTE is the flattest and the worst: preservation sits near 50.6% in *every* target language (Figure 1). Quotation-mark integrity is essentially independent of the target language, which suggests a decoder-level rather than a language-specific cause.

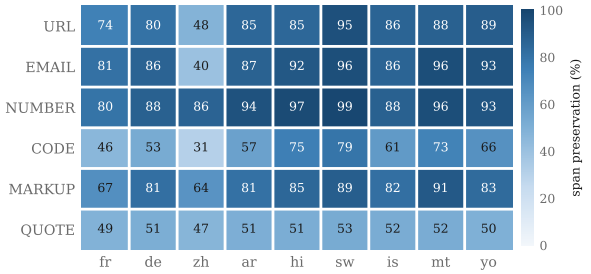


Figure 1: Span preservation (%) by span type and target language, NLLB-200-distilled-600M, 73,881 spans. Darker is better.

Language dependence is large, but the direction is the opposite of what we predicted. Preservation ranges 18.7 points, from 76.7% (fr) to 95.4% (sw). The ordering, however, *inverts* H2: the high-resource targets are damaged most (fr 76.7%, zh 79.6%, de 84.3%) and the lower-resource ones least (sw 95.4%, hi 93.4%, mt 92.7%). **H2 as pre-registered is refuted, with the sign reversed.**

The obvious confound is non-translation: if the model simply copies English through for low-resource targets, their spans survive trivially. Copy-through is indeed about twice as common there (32.5% for sw against 15.0% for fr and 7.0% for zh). Restricting to units the model actually rewrote lowers every language by 2–5 points and leaves the ordering intact (fr 73.0%, zh 78.0%, sw 93.4%), so the inversion is a property of translation behaviour, not of its absence. Chinese is the limiting case: it has the lowest copy-through (7.0%) — it rewrites most aggressively — and destroys URLs, emails and code worst of all. Span damage tracks how much the system rewrites.

Failures are dominated by deletion, not transliteration. Of 81,512 failures, 46.8% are outright *deletion* and 18.0% truncation or splitting. Script transliteration — the failure mode a digit-system argument would predict, and the one most visible in prior controlled studies — accounts for just 0.3% overall (5.0% for Hindi, 0.3% for Arabic). Spans are overwhelmingly not converted; they are lost.

A further 6.6% are delimiter substitutions, many of which are locale-*correct* typography ("…" to «…»). The same caveat applies to numbers: French NUMBER scores 79.6% under the strict matcher but 90.4% once separators and numeral

systems are ignored, because `14,000` becomes `14 000`. This is good translation and unusable data, and we report both brackets rather than choosing one.

Round-trip similarity does not predict span damage (H3). Round-trip chrF2 averages 65.7 and spans 20.8 points across languages, but it barely tracks preservation. Across the nine language means, Spearman $\rho = 0.48$ ($p = 0.19$, $n = 9$): not significant, and below the $|\rho| \geq 0.6$ threshold we fixed for treating fluency as a usable predictor. Per unit, over 111,573 scored segments, $\rho = 0.25$ — certain in sign given the sample size, but explaining roughly 6% of variance. Masking spans before scoring, which removes the mechanical coupling whereby damaged spans depress round-trip chrF2, lowers it further to $\rho = 0.21$. **H3 is supported**: fluency is detectably but negligibly informative about span damage. Figure 2 shows the dispersion: Arabic and French have near-identical chrF2 yet differ by five points of preservation.

Downstream cost. Of the spans that were syntactically valid before translation, 87.0% remain recoverable and valid afterwards. For code the figure is 59.8% and for quotations 48.3%. A pipeline that re-extracts URLs, addresses or code from an MT-built corpus should expect to lose roughly one span in eight, and far more for the structured types.

Second system: the effect replicates, the ranking does not. On the 10% subsample, Llama-3.3-70B preserves 82.6% of spans overall — the same order of damage as NLLB, with much lower copy-through. The per-language profile, however, differs sharply: it is markedly better on Chinese (88.4% against 79.6%) and markedly worse on Hindi (68.1% against 93.4%). Structured-span damage is therefore a property of MT-based corpus construction rather than of one model, but *which* languages suffer is system-specific and cannot be read off a published table.

5 Conclusions and recommendations

Report span preservation next to the quality score. Span preservation is cheap to compute, needs no reference translation and no model, and measures what aggregate surface metrics cannot see. Any release of an MT-built corpus should report a preservation rate per span type and target language beside its quality figure; our detectors and scoring code are released for that purpose.

	NLLB-200			Llama
	Preserv.	Transl. only	chrF2	Preserv.
fr	76.7	73.0	68.1	73.5
de	84.3	79.8	68.7	80.6
zh	79.6	78.0	53.8	88.4
ar	89.9	87.9	61.6	81.3
hi	93.4	91.7	70.8	68.1
sw	95.4	93.4	74.5	89.0
is	84.9	80.6	64.2	83.7
mt	92.7	89.5	66.7	91.2
yo	89.7	84.9	62.7	87.3

Table 1: Per-language span preservation (%) for both systems, against round-trip fluency. "Transl. only" excludes units the model returned unchanged. Higher-resource targets are damaged *more*, and fluency does not order the columns.

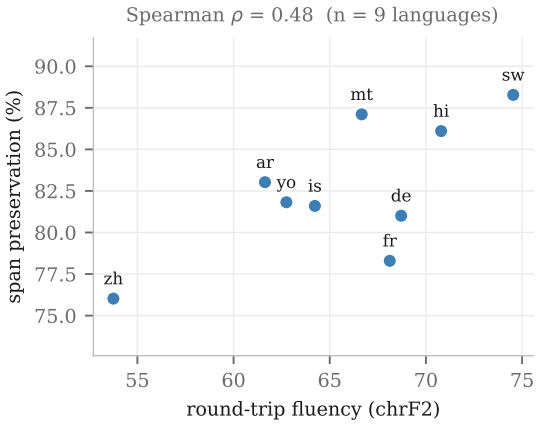


Figure 2: Round-trip fluency does not track span preservation across target languages.

Protect high-value spans rather than repair them. Because our segmenter never splits a span, the damage happens inside the translation model, not around it, so the remedy is not post-hoc repair but the placeholder discipline standard in commercial localisation and largely absent from corpus building: mask detected spans, translate, restore. It also converts an unbounded problem into a bounded one—detector coverage—which can be measured (Appendix A).

State per-language caveats, especially for low-resource targets. A single corpus-level quality claim hides the variation we observe across target languages—and the languages with the least native web text are exactly those most likely to be built by translation, so the caveat matters most where it is least often given.

Is the damage necessarily harmful? A reviewer raised the reasonable counter-hypothesis that losing URLs and email addresses is a side benefit,

scrubbing personal information from training data. The question separates by span type. For EMAIL the point has force, though corruption is a poor substitute for a filter that can be audited. For the types carrying most of the mass it does not: NUMBER is 62,157 of our 73,881 spans, and a silently altered quantity, version string or identifier is not anonymised but *wrong*—undetectably so, because the segment still reads well. The useful distinction is removal, which is visible and can be a policy choice, against alteration, which is neither: 46.8% of failures are deletions, but 3.8% are digit alterations and 22.7% other alterations.

Distinguish reformatting from destruction. Not every non-preserved span is an error: locale-correct rewriting—`14,000` to `14 000`, `"..."` to «...»—is good translation and bad exact-match data. Reporting the strict rate with a lenient bracket (digit-sequence recovery) separates the two and keeps the headline number honest.

Conclusion. Translating English web text to build a corpus in another language damages 12.6% of the structured spans it contains, and the corpus-level scores used to validate such corpora do not reveal it. The damage is not uniform: it concentrates by span type, from 91.1% preservation for numbers down to 50.6% for quotation delimiters, and it varies by target language in the direction opposite to the one we pre-registered. None of this argues against building corpora by translation; it argues that span preservation belongs next to the quality score when such a corpus is released, and that masking spans before translation is the cheap fix. We release the annotated corpus, the detectors and a verifier so both can be checked.

Limitations

Systems and settings. We test one dedicated NMT model and one instruction-tuned LLM, at fixed deterministic decoding settings. Larger NLLB variants, other architectures, and sampling-based decoding may behave differently. The second system runs on a 10% subsample and is a robustness check, not a co-headline.

English source only. All source text is English. Translation out of other source languages, and pivot translation, are untested; the span inventory of non-English web text also differs.

Detector coverage. Our detectors are frozen, documented and measured (precision 98.5%, recall 91.4% on a held-out development split), but they are regular expressions, not an oracle. Straight single-quoted strings, numeric tokens under three characters, and digit runs followed by letters are excluded by definition; version strings and hyphenated package filenames are known recall gaps. Missed spans do not enter the denominator, so coverage limits what we can see rather than biasing the rates we report.

WET extraction. WET is plain text, so HTML structure is gone before we see it. "Markup" here therefore means markup appearing as content — entities and literal tags in tutorials and documentation — and MARKUP is much the rarest of our six types (359 spans), with correspondingly wide intervals. Measuring page scaffolding would require WARC and an extraction step that would itself confound the comparison.

Segment-level translation. Units are segments, not documents. This matches how sentence-level NMT is actually applied and makes our numbers identical to full-corpus translation for those segments, but it means we cannot observe document-context effects.

Parse proxy is syntactic. We test whether a span is recoverable and well-formed after translation, not whether a URL still resolves or an address still receives mail. Live resolution would be neither reproducible nor kind to the hosts involved.

Ethics Statement

The source text is derived from Common Crawl and may contain personal data such as the email addresses our EMAIL detector is designed to find. We release span annotations and verdicts together with the document URLs and WET file identifiers needed to reconstruct the sample, and we do not construct or distribute any aggregated contact list. A random web sample also contains adult and spam content; we apply no topical filtering, since filtering on content would bias the span inventory we set out to measure.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a

collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *American Journal of Computational Linguistics*, 29(3):333–348.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 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, 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*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

A Detector quality

Span type	Precision	Recall
URL	95.0	100.0
EMAIL	100.0	100.0
NUMBER	100.0	92.0
CODE	97.5	71.4
MARKUP	100.0	–
QUOTE	100.0	100.0
Overall	98.5	91.4

Table 2: Detector quality (%), hand-adjudicated on the development split (204 sampled spans; 96 detected and 9 missed spans in 25 exhaustively marked windows).

Counting Documents Is Not Counting Text: Unit Bias in Web-PDF Corpus Statistics

Luca Foppiano
Common Crawl Foundation
luca@commoncrawl.org

Abstract

PDF corpora advertise their size in tokens, but every rate they publish (coverage, OCR routing, re-fetch recovery, language mix) is computed per document, and none decomposes its token total. Because PDF length is extremely skewed, the two units can describe the same corpus very differently. We ask how the headline statistics of a web-PDF corpus change when each document is weighted by the text it contributes rather than counted once. We used `CC-MAIN-2021-31-PDF-UNTRUNCATED` (7.9M Common Crawl PDFs, 32.6B tokens), the one public corpus that pairs the fragments Common Crawl stored with the re-fetched originals. Text mass is highly concentrated: 3.02% of text-bearing documents hold half the tokens (Gini 0.807). The clearest consequence is Common Crawl's payload cap, which truncated 23.06% of these documents but 63.08% of their text. Reconstructing the truncated fragments and extracting both versions, two widely used text-layer parsers recover only 1.4% and 11.4% of that exposed text, so roughly 55–62% of the corpus's text is unrecoverable from the crawl by such pipelines; under the 5 MiB cap adopted in March 2025, 30.19% of tokens would still be exposed. We recommend that corpus statistics be reported in both units, documents and tokens.

1 Introduction

PDFs have become a first-class source of pretraining text (Poznanski et al., 2025; Montalvo and Wightman, 2024). In one current open data pool, PDF-derived text amounts to roughly 4.3T tokens (Marin Community, 2025).[1] FinePDFs alone is described as "about 3 trillion tokens across 475 million documents" (Kydlíček et al., 2025). How do a web-PDF corpus's headline statistics change when each document is weighted by the text it contributes rather than counted once?

Those corpora are not naive about tokens: they headline them. What none of them does is compute a *rate* in tokens, or break a token total down. FinePDFs routes 368.8M of 1.29B *files* to OCR and recovers 53.5% of its truncated *files* by re-fetching (Kydlíček et al., 2025); CCpdf reports per-*document* success rates (Turski et al., 2023); PDFA gives three totals (2,159,432 documents, 18M pages, 9.7B tokens) but no drop rates in any unit (Montalvo and Wightman, 2024). No PDF corpus we are aware of decomposes its token total by length or quality.

This would be harmless if documents were interchangeable. They are not. The web-text community already recognises this: Nemotron-CC reports a *token*-weighted yield, "+57.4% more high-quality tokens" (Su et al., 2025), and PDFs have far more extreme length skew than HTML. FinePDFs itself observes that its documents have a "∼5.3k character median (about 2× other corpora)" with a "95th percentile ∼68k" against "∼11–13k elsewhere" (Kydlíček et al., 2025), and then reports every rate per document anyway.

We measure what those rates become when weighted by text. The shipped metadata of `CC-MAIN-2021-31-PDF-UNTRUNCATED` lets us reweight three of the quantities above, document length, producer toolchain and truncation; the argument applies equally to OCR routing, filter drop and language mix, which we do not measure here. Our contributions are:

1. the first token-weighted characterisation of a web-PDF corpus, with a direct document-versus-token comparison of every headline statistic (§4.1–4.3);
2. a concentration result: text mass in web PDFs is far more skewed than document counts suggest (Gini 0.807);

[1] Marin's `datakit` source registry: FinePDFs 1,186.5B English plus 1,353.4B across 18 non-English subsets, and `dolma4pdfs` 1,804.0B over 137,132,279 documents. https://github.com/marin-community/marin

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 107–114
October 29, 2026 ©2026 Association for Computational Linguistics

3. the first measurement of what Common Crawl's truncation cap removes in text: 23.06% of documents but 63.08% of tokens, of which two text-layer parsers recover 1.4–11.4%, with 30.19% still exposed at the new 5 MiB limit (§4.4–4.6).

2 Related Work

Totals in tokens, rates in documents. FinePDFs (Kydlíček et al., 2025) routes documents between a text path (Docling) and a GPU OCR path with a learned classifier reporting F1 0.71 on the OCR class, and quantifies every stage in files. CCpdf (Turski et al., 2023) tabulates "number of documents per processing step and language" and reports no token count for its own corpus. PDFA (Montalvo and Wightman, 2024), derived from the same corpus we study, filters per document, discarding files over 100 MB or slower than 500 ms to render, i.e. selecting on *size*, and publishes no drop rates. GovScape (Huang et al., 2026) excludes "pages with no text embedded within the PDF" without quantifying the exclusion in any unit. Extraction benchmarks (Ouyang et al., 2025; Poznanski et al., 2025) measure fidelity on documents that parsed and report no coverage at all. Kuznetsov and Gurevych (2023) argue more generally that what counts as "text" is under-specified in NLP. In language-model research the question has been settled in practice rather than in principle: the token is the unit in which corpora are sized, data mixes are set and compute is scaled, and it is the unit a model consumes. Corpus statistics have not followed consistently.

Truncation is named but never quantified. Common Crawl caps the payload it stores per record. PDFs are the format worst affected: for the crawl `CC-MAIN-2023-06` they were 0.8% of successfully fetched records but 11.85% of 88 TiB of WARC storage (PDF Association, 2023). CCpdf stated the consequence plainly in 2023: the crawler's 1 MB cap is "quite high for HTML pages, but unfortunately rather low for PDF files" (Turski et al., 2023). Their workaround was to re-download the truncated documents from origin; to our knowledge nobody since has quantified the effect.

From `CC-MAIN-2025-13` (released in March 2025) the cap rose from 1 MiB to 5 MiB, reported as +13% fetched content (403 TiB → 455 TiB) (Common Crawl Foundation, 2025).

The Common Crawl Foundation reported the effect of the change on truncation by MIME type: PDFs fell from 25.7% to 6.8%, HTML from 2.2% to 0.04%. On average, all types fell from 2.25% to 0.14%. Even after the change PDFs are truncated at roughly 49× the rate of content generally. What is not reported, before or after, is how much text that removes.

3 Data and Method

Corpus. `CC-MAIN-2021-31-PDF-UNTRUNCATED` (Digital Corpora et al., 2021) contains every PDF found in Common Crawl `CC-MAIN-2021-31`, with payloads that Common Crawl truncated *refetched whole* from origin. It ships five metadata tables covering 8.3M URLs. Because `file_name` is the post-SHA-256 identity, we deduplicate to one row per file, giving 7,932,654 unique documents.

The unit. Text mass is taken from `tika_eval_num_tokens`, a token count already published as part of the corpus metadata (Digital Corpora et al., 2021) (Apache Tika 2.8.0, tesseract disabled), a whitespace/ICU count rather than an LLM tokenizer's. The total is 32,570,135,761 tokens. We claim only proportions, which is all the argument requires. Tika stops counting at 10,000,000 tokens, so a longer document is recorded at that value; only two documents in the corpus reach it, and both are therefore undercounted.

Measures. For any subset S of the N documents we report its *document share* $|S|/N$ and its *token share* $\sum_{d\in S} t_d / \sum_d t_d$, where t_d is the token count of document d; the "ratio" column of Table 1 is the second divided by the first. The *head share* at level q is the smallest fraction of documents, taken in decreasing order of t_d, whose tokens reach a fraction q of the total. Concentration is summarised by the Gini coefficient in its standard form: with the n documents sorted so that $t_{(1)} \le \cdots \le t_{(n)}$,

$$G = \frac{2\sum_{i=1}^{n} i\, t_{(i)}}{n\sum_{i=1}^{n} t_{(i)}} - \frac{n+1}{n}. \qquad (1)$$

$G = 0$ means every document carries the same amount of text and $G \to 1$ means all of it sits in one document; values around 0.3–0.4 are typical of national income distributions, the setting the coefficient comes from. We compute it over the $n = 7{,}292{,}093$ documents with at least one token, so the 8.1% of zero-token documents do not inflate it, and the two censored documents can only lower

it. Figure 1 shows the underlying Lorenz curve. The categories of Table 1 are bins on page count and first-page orientation, chosen so that the two extremes, single-page files and files over 50 pages, are unambiguous; they are not genre labels, and a long form is binned with long reports.

Exposure, recovery, loss. Three quantities recur in §4.4–4.6. *Exposure* E is the token share of the documents Common Crawl truncated; it is computed from the shipped metadata over all 7.9M documents. *Recovery* $r_{e,d}$ of a truncated document d under extractor e is the number of tokens e extracts from the truncated document, divided by the number it extracts from the whole document. The corpus rate r_e sums both quantities over the documents processed before dividing, so long documents weigh more. It is measured by re-creating the fragments (§4.5). *Loss* is $E\,(1 - r_e)$: the share of the corpus's text that a pipeline reading the text layer with e cannot obtain from the crawl. Loss is an estimate under two assumptions, both tested in §4.5: that recovery on the documents we processed transfers to the exposed mass as a whole, and that a document's weight is the same whether counted in Tika tokens or in the whitespace tokens of our own extraction.

Truncation. The provenance table records `cc_truncated`, `fetched_status`, `fetched_length` and the WARC byte range per file (`cc_truncated='length'` coincides exactly with `fetched_status='REFETCHED_SUCCESS'`). For truncated records the WARC record length pins to ∼1.049 MB, so the cap is directly observed, and the re-fetched originals give the true size of every document Common Crawl stored only as a fragment: a pairing no other public corpus offers.

4 Results

4.1 Text mass is extremely concentrated

Table 1 gives the page-count distribution in both units. Counting by page count alone, regardless of orientation (memo row of Table 1), documents over 50 pages are 5.00% of the corpus and 53.53% of its text: a document in that tail carries about 10.7× an average document's text. We use this threshold for the concentration claim here and in §4.2. The mirror image, from the two bottom rows above the rule: the 46.2% of documents with three pages or fewer contribute 6.46% of the text. Over the 7,292,093 text-bearing documents, 3.02% hold half the tokens and 15.54% hold 80% (Figure 1); the Gini coefficient of tokens across documents (Eq. 1) is 0.807.

Bin	doc %	tok %	ratio
>50 pp	4.42	**49.71**	**11.24**×
4–30 pp	31.64	26.87	0.85×
31–50 pp	3.21	9.08	2.83×
landscape, any length	14.08	7.76	0.55×
2–3 pp	21.20	3.76	0.18×
1 pp	24.96	2.70	**0.11**×
>50 pp, any orientation	5.00	53.53	10.71×

Table 1: The same corpus counted two ways. Rows above the rule are mutually exclusive bins on page count and first-page orientation; orientation takes precedence, so a long landscape document falls in the landscape row and the page-count rows cover portrait documents only. The memo row below the rule ignores orientation and is the threshold used in the text and in §4.2.
0.49% of documents with unparseable page metadata are omitted.

Counted by document, the corpus is dominated by files of three pages or fewer; counted by text, it is a corpus of long documents. Both descriptions are arithmetically correct.

4.2 The concentration is not an artifact of one extractor

The concentration in §4.1 rests on one extractor and one way of counting: Tika's text extraction and its whitespace token count. Either could favour long documents, the first if short files are more often scans or forms that yield little text, the second because the count depends on script and layout. We therefore validate the distribution against page counts, which the corpus authors extracted with Poppler, an independent tool measuring an entirely different quantity. Page counts reproduce the same concentration: the Gini coefficient of *pages* across documents is 0.767 against 0.807 for tokens, 3.98% of documents hold half the pages against 3.02% for tokens, and documents over 50 pages account for 54.27% of all pages against 53.53% of all tokens, a difference of 0.74 percentage points. Two tools measuring two different quantities give the same answer: a shared extraction bias cannot produce this agreement.

4.3 The TEX share more than doubles

Classifying the `producer/creator` strings into toolchain families, a TEX toolchain accounts for 1.66% of documents but 4.05% of tokens (2.43×). PowerPoint runs the other way, 2.24% of docu-

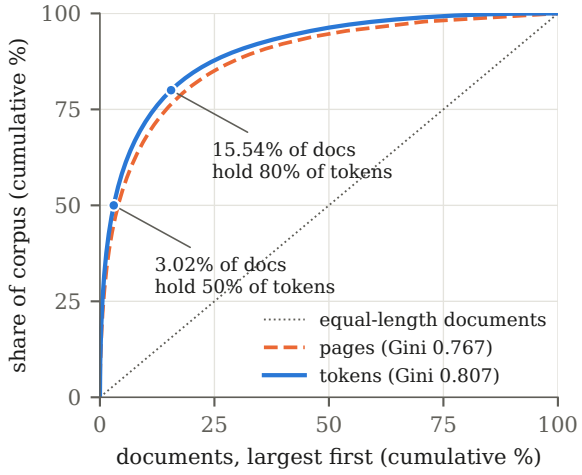


Figure 1: Cumulative share of tokens (solid) and of pages (dashed) held by the largest documents, over the 7,292,093 text-bearing documents and the documents with a page count respectively. The marked points are the head shares quoted in the text: 3.02% of documents hold 50% of the tokens and 15.54% hold 80%. The dotted diagonal is a corpus in which every document has the same length.

	docs %	**tokens %**
Truncated by CC (1 MiB)	23.06	**63.08**
still truncated at 5 MiB	5.94	**30.19**

Table 2: Common Crawl's truncation cap, expressed in both units, over the 7,932,878 files with provenance records.

ments against 0.82% of tokens (0.37×), and Microsoft Word is 24.95% of documents against 19.37% of tokens.

Toolchain is a proxy for the production pipeline, not for genre: Word and PowerPoint also carry scholarly text, and producer strings are self-reported and incomplete (see Limitations). What the comparison shows is that the TEX share differs by 2.4× between the two units, and the token unit is the one a language model consumes.

4.4 Truncation: 23% of documents, 63% of the text

This section measures exposure, the token share of the documents Common Crawl truncated; §4.5 measures how much of it is recoverable.

Common Crawl truncated 1,829,061 of this corpus's documents at the 1 MiB cap in force at the time, corresponding to 23.06% of them. Those documents hold 63.08% of the corpus's text (Table 2). Truncated documents carry 2.74× the mean token mass; their median true size is 2.62 MB against 190 KB for the rest.

Applying the current 5 MiB cap to the corpus, 5.94% of documents and 30.19% of tokens would still be truncated. The March 2025 change (§2) therefore removes roughly half of the exposure and leaves the other half in place. Our document-level counterfactual is consistent with Common Crawl's own post-change measurement of 6.8% on 2025 crawls (Common Crawl Foundation, 2025), computed on a population five years younger; the token figure has no published counterpart.

In bytes, the 1 MiB cap kept 1.738 TiB of these files' 9.449 TiB total, discarding 81.6%; storing them whole would have cost 7.712 TiB of additional archive. The 5 MiB cap lands between the two: it would keep 5.106 TiB (46.0% discarded), so the March 2025 change spends 3.368 TiB of that 7.712 TiB and leaves 4.344 TiB behind. At either cap, byte loss overstates text loss: 81.6% of bytes against 63.08% of tokens at 1 MiB, 46.0% against 30.19% at 5 MiB, because PDF bytes are largely images and embedded fonts. The byte figure measures the archive; the token figure measures the corpus.

4.5 Little of the exposed text is recoverable

This section measures recovery, as defined in §3, and combines it with the exposure of §4.4 into loss. To do so, we reconstruct what Common Crawl held, cutting each of this corpus's 1,829,061 truncated documents at 1 MiB, and extract text from both the fragment and the whole file. Whether a fragment is recoverable may be a property of the parser rather than of the file, so we run two independent engines, PyMuPDF (MuPDF) (Artifex Software, Inc., 2025) and PDFium (Chromium's) (pypdfium2 team, 2025), and report on the 1,225,130 documents both tools successfully processed, and whose whole version yields text.

PyMuPDF recovers 11.4% of the tokens and 20.7% of the pages; PDFium recovers 1.4% and 1.9%, a gap of 8.4× on identical input. The two agree to within 1.7% on the *intact* versions of those same documents, so the divergence is specific to damaged input and is not a general difference in extraction quality. There is therefore no single "recovery rate" for a truncated PDF: the quantity is not defined until the extractor is named.

The shape of the failure differs as sharply as its size, and both shapes are invisible to a pipeline that counts successful parses. PyMuPDF opens 98.9% of truncated files, because MuPDF rebuilds

a missing cross-reference table, but the page content streams lie beyond the cut, so the document opens, reports a page count, and returns no text: 72.4% yield nothing at all. PDFium attempts no such reconstruction and refuses the file outright, failing to open 96.7% of the same documents; its zero-yield rate of 96.8% is almost exactly its open-failure rate. One extractor logs these as successes with empty output, the other as parse errors. The gap between the engines is concentrated in the 59.3% of truncated documents that are not linearized; Appendix B breaks recovery down by linearization.

Applying each engine's recovery to the exposure in Table 2, the loss $E\,(1 - r_e)$ is $63.08\% \times (1 - 0.114) \approx 56\%$ of the corpus's text under PyMuPDF and $63.08\% \times (1 - 0.014) \approx 62\%$ under PDFium; we report the range 55–62%. The product multiplies an exposure counted in Tika tokens by a recovery counted in whitespace tokens. Since every truncated document has both counts, we reweight each by its Tika count instead. The corpus figure moves by 0.42 points under PyMuPDF (to 55.3%, the lower end of the range) and by zero under PDFium. The two counts agree closely per document (median ratio 1.01–1.02). They diverge only for writing systems without whitespace word separators, such as Japanese and Chinese, which is a property of the corpus rather than of truncation. OCR is not an alternative route: a page can be rasterised only if its page object and content stream lie inside the fragment, in which case its text layer is extractable as well, and Poppler, the renderer behind the usual `pdf2image` OCR front end, refuses truncated files as PDFium does. Refetching from origin, as CCpdf and FinePDFs did, is a remedy for the crawl rather than an extraction method, and it is what §4.6 argues for. One caveat does remain: we count tokens rather than reading them, and a token count cannot distinguish text that a repair path recovered cleanly from reconstruction artefact, so "recovers more" must not be read as "recovers better".

4.6 The 5 MiB cap recovers little of what it still exposes

Raising the cap to 5 MiB (Table 2) leaves 30.19% of tokens *exposed*; we tested whether they are *recoverable*, since every document truncated at 5 MiB is also truncated at 1 MiB and only the cut point moves between the runs.

On the 316,174 documents PyMuPDF processed at both caps, recovery rises from 3.3% to 13.2% of tokens, and 21,203 documents (6.7%) go from yielding no text at all to yielding some. PDFium moves from 0.1% to 2.4% on the same documents. A five-fold larger prefix therefore multiplies recovered text roughly four-fold, but still leaves 77.6% of these documents (97.0% under PDFium) yielding no text. The change halves the exposure; what it leaves behind remains almost entirely unreadable.

The benefit is also concentrated immediately above the cap (Table 3, Appendix A): 74% of the rescued documents lie between 5 and 10 MiB, while above 25 MiB the larger cap is worth 2–3 percentage points. This is what a fixed prefix must do, 5 MiB is half of a 10 MiB file and a twentieth of a 100 MiB one, and it holds under both engines. Raising the cap further has sharply diminishing returns per additional TiB of archive: the documents that dominate the remaining token mass are those a larger fixed prefix helps least.

Both effects reflect the cut rather than the run: token counts from the whole documents agree exactly for all but 0.036% of paired documents under both engines (a shared, harness-imposed deadline). Recovery is not monotonic, however: 800 documents (0.25%) yield text at 1 MiB and none at 5 MiB, so repair is sensitive to *where* a file stops, but the effect is 26× rarer than the reverse and does not disturb the aggregate.

5 Conclusion

Counted in documents and counted in text, the studied corpus `CC-MAIN-2021-31-PDF-UNTRUNCATED` is two different corpora: the units diverge by up to 11× per category, and Common Crawl's cap affected 23% of documents but left 55–62% of the text unrecoverable to text-layer parsers. Corpus statistics should therefore be reported in both units; every rate in this literature (coverage, filter drop, OCR routing) is per-document or unreported. Any filter that selects on size, such as PDFA's 100 MB and 500 ms cuts, FinePDFs' OCR router or the crawler's cap, removes far more text than documents, because size is where the text is; its effect should be reported in tokens. For Common Crawl, §4.4–4.6 quantify the 1 MiB → 5 MiB change: it halves the exposure, but little of what it still truncates is recoverable; the remaining exposure calls for a size-aware fetch policy for large PDFs, not a higher cap.

Limitations

Every figure describing corpus *composition* and truncation *exposure* is computed over all 7.9M documents from the shipped metadata. The recovery figures cover the 1.2M truncated documents both extractors successfully processed, 67% of the 1.83M Common Crawl truncated. The shortfall is concentrated in ten of our sixty-four PyMuPDF shards, which produced nothing because of technical errors in the extractor (e.g. out of memory, hanging); since shards partition the corpus by archive index, the loss is arbitrary with respect to document content, and per-shard recovery rates vary by about a percentage point. It is nonetheless not a random sample, and we report no confidence intervals. The metadata is third-party and derived (Digital Corpora et al., 2021), and `producer`/`creator` strings are self-reported with 8.08% unclassified, so the TEX share is a lower bound in both units. The page-count bins of Table 1 and the toolchain families are heuristics, not genre labels: a long tax form falls in the >50-page bin, and a thesis typeset in Word falls outside the TEX family. `page_size` is recorded for the first page only, making the landscape category approximate. We study a single crawl snapshot, from before the policy change, and the only one for which re-fetched originals exist; the 5 MiB figure is therefore a counterfactual computed on 2021 file sizes rather than a measurement of a current crawl. Re-measuring exposure on a post-2025 crawl needs only the WARC metadata and is the obvious next check; we have not done it. Finally, both extractors are text-layer parsers and we run no OCR arm. A page can only be rendered if its objects lie inside the truncated document, in which case its text layer is extractable as well, so OCR cannot recover more from the crawl than a text-layer parser (§4.5).

Acknowledgement

We thank DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH) for supporting this work with computation and storage.

References

Artifex Software, Inc. 2025. PyMuPDF. Version 1.28.2; Python bindings for MuPDF.

Common Crawl Foundation. 2025. March 2025 crawl archive now available. https://commoncrawl.org/blog/march-2025-crawl-archive-now-available. Truncation limit raised from 1 MiB to 5 MiB at CC-MAIN-2025-13.

Digital Corpora, NASA JPL, and DARPA SafeDocs. 2021. CC-MAIN-2021-31-PDF-UNTRUNCATED (SAFEDOCS). https://digitalcorpora.org/corpora/file-corpora/cc-main-2021-31-pdf-untruncated/.

Ying-Hsiang Huang, Claire Gong, Shreya Shaji, Alison Yan, Leslie Harka, Albert Du, Anjali Gopal, Samuel J Klein, Shannon Zejiang Shen, Mark Phillips, Trevor Owens, Kyle Deeds, and Benjamin Charles Germain Lee. 2026. Govscape: A public multimodal search system for 70 million pages of government pdfs. *Preprint*, arXiv:2511.11010. ArXiv v3; quotation from Sec. 2.2.

Ilia Kuznetsov and Iryna Gurevych. 2023. An inclusive notion of text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics. ArXiv:2211.05604.

Hynek Kydlíček, Guilherme Penedo, and Leandro Von Werra. 2025. Finepdfs: Liberating 3t of the finest tokens from pdfs. https://huggingface.co/datasets/HuggingFaceFW/finepdfs.

Marin Community. 2025. Marin: an open lab for building foundation models (`datakit` source registry). https://github.com/marin-community/marin. Accessed September 2026.

Pablo Montalvo and Ross Wightman. 2024. PDFA: pdfa-eng-wds. https://huggingface.co/datasets/pixparse/pdfa-eng-wds.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24838–24848. IEEE.

PDF Association. 2023. New large-scale pdf corpus now publicly available. https://pdfa.org/new-large-scale-pdf-corpus-now-publicly-available. Reporting WARC storage and truncation shares for CC-MAIN-2023-06.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*.

pypdfium2 team. 2025. pypdfium2. Version 5.12.1; Python bindings for PDFium.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 2459–2475.

Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. 2023. CCpdf: Building a high quality corpus for visually rich documents from web crawl data. In *Document Analysis and Recognition – ICDAR 2023*. ArXiv:2304.14953.

A Recovery by document size

Table 3 breaks the paired recovery measurement of §4.6 down by document size.

		recovered		
size	docs	1 MiB	5 MiB	gain
5–10 MiB	179,810	4.88%	22.44%	+17.6
10–25 MiB	99,831	2.62%	8.97%	+6.4
25–50 MiB	25,700	1.64%	4.47%	+2.8
50–100 MiB	8,683	1.36%	3.28%	+1.9
>100 MiB	2,150	1.01%	2.96%	+2.0

Table 3: What raising the cap buys, by document size (PyMuPDF, gain in percentage points). The same 316,174 documents are cut at both points; the benefit is concentrated immediately above the cap.

B Recovery by linearization

The gap is not spread evenly across the corpus. A *linearized* PDF (Adobe "Fast Web View") places its first page and a cross-reference table at the front of the file so that a partial copy is still renderable; 40.7% of the documents here are linearized. Under PyMuPDF those files recover 1.1% of their tokens against 18.5% for the non-linearized rest, while under PDFium it is 1.1% against 1.5%. Note that the gap persists when files of similar size are compared, so it is not a size effect. The difference between the two engines therefore lives in the non-linearized population, and the Fast Web View structure, designed to keep partial files usable, is the one from which least is recovered.

Is educational content enough? The linguistic and thematic effects of educational filtering

Urho Kunnas, Veronika Laippala, Sampo Pyysalo, Tomasz Galica
TurkuNLP, University of Turku
{urho.r.kunnas, mavela, spyysalo, tomasz.galica}@utu.fi

Abstract

Corpus filtering strategies have evolved steadily, leading to increasingly effective data selection techniques that improve downstream LLM performance. However, while most filters focus on knowledge-rich and educational content, their effect on the linguistic and thematic composition of the documents in pre-training corpora remains underexplored. In this paper, we investigate how educational filtering affects both linguistic and thematic content by combining Propella-1 educational value annotations and the FinePDFs-Edu filter with the MultiCORE register (genre) framework. Focusing on the Finnish subset of the HPLT 4.0 corpus, we examine how educational filtering changes the distribution of registers and how educational value relates to register categories. By studying machine translated documents with high educational value we show that educational filters focus on content over textual quality. We further analyze changes in the thematic composition using Propella-1 business sector labels, examining whether educational filtering shifts the representation of different content domains. Finally, to assess the impact of filtering, we trained three small language models and evaluated their downstream performance on benchmark tasks. We find that educational filters disagree in their assessment of document educational value, show similar preferences for linguistic registers, but differ in the thematic content they select, resulting in different downstream performance across task types.

1 Introduction

Web-corpora are the main source of pre-training data for large language models (LLMs) such as Apertus (Hernández-Cano et al., 2026), Olmo3 (Olmo et al., 2026) or DCLM-7B (Li et al., 2025). As the amount of web data continues to grow, data curation objectives have shifted from solely maximizing data volume towards improving average data quality. Higher-quality pre-training data enables LLMs to achieve the same or better downstream performance using fewer training tokens. This in turn decreases LLM training cost and time (Li et al., 2025).

A popular strategy for improving the quality of pre-training corpora is the use of educational filters offering either binary or ordinal classification. These filters estimate the educational value of documents, defined as their usefulness for learning and knowledge transfer. Documents containing clear explanations, structured information, instructional or factual content are graded with higher educational value than e.g. entertainment-oriented text.

Studies on educational filters demonstrate that higher-quality data can substantially improve training efficiency. Models trained on FineWeb-Edu (Lozhkov et al., 2024) and Ultra-FineWeb (Wang et al., 2025) matched or exceeded the English downstream benchmark performance of a model trained on the original FineWeb dataset (Penedo et al., 2024) while using less training data. However, these gains come at the cost of aggressive filtering: FineWeb-Edu removes 92% and Ultra-FineWeb 93% of the data originally available in FineWeb.

Such substantial reductions may change the distribution of pre-training corpora by reducing topic coverage and removing cultural references or other information that is not directly captured by standard benchmarks. Moreover, relying on downstream performance is challenging, as such benchmarks capture only a subset of language model capabilities and do not directly assess how changes in training data composition affect the model's coverage of different knowledge domains, language varieties, or text types. Furthermore, it remains unclear how content-based filters such as FineWeb-Edu respond to stylistic and linguistic features of texts. This is particularly important given the many dimensions of linguistic variation that may be overlooked when filtering primarily for educational value, including

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 115–130
October 29, 2026 ©2026 Association for Computational Linguistics

dialects, informal language, and situational variation.

Our linguistic analysis focuses on situational variation using the Multilingual CORE (MultiCORE) register (genre) framework (Henriksson et al., 2024) targeting the entire multilingual web. The framework categorizes web documents according to *registers* (Biber and Conrad, 2019), situationally defined text varieties such as news articles, discussion forums and how-to pages. Registers cover systematic differences in the situational context of the text, including communicative purpose, writing style, and linguistic characteristics, and have been shown to be one of the strongest predictors of linguistic variation (Biber, 2012).

We analyze how educational filtering changes the distribution of registers and thematic content in the Finnish subset of HPLT 4.0. In our study, we employ the FinePDFs-Edu classifier-based quality filter (CQF) (Lozhkov et al., 2024) and Propella-1-4b annotations (Idahl et al., 2026) for educational value and business sector labels.

FinePDFs-Edu is a well established CQF, trained on 1.3M annotations generated by Qwen3-235B-A22B-Instruct-2507 (Team, 2025) for FinePDFs (Kydlíček et al., 2025) dataset samples. FinePDFs-Edu assigns continuous scores, where higher scores indicate a better educational value of the document.

Propella-1 is a family of multilingual models that annotate documents across six semantic dimensions. In this work, we focus on two of the dimensions: educational value and business sector. The educational value annotation enables a direct comparison with the FinePDFs-Edu filter, while business sector labels are used to analyze how educational filtering affects the thematic composition of the corpus. Business sector categories include, for example, academic research, human resources, media and entertainment, and the gaming industry.

To investigate how educational filters and associated changes of register and thematic content distribution affect model behavior, we trained three small language models. We use model trained on unfiltered Finnish subset of HPLT 4.0 as a reference model, and compare it with models trained on data filtered by FinePDFs-Edu and Propella-1.

2 Related work

Data filtering strategies have been discussed in the past in multiple contexts beyond purely improving the quality of pre-training corpora. Albalak et al. (2024) show two main approaches to data filtering: distribution matching and distribution diversification, with classifier based quality filtering (CQF) typically belonging to the former category. Albalak et al. (2024) shows that CQF can introduce bias in processed datasets: against certain groups, demographics, dialects, or data types. Additionally, one-dimensional quality classifiers such as educational filters can unintentionally favor stylistic features over semantic content. For example, Klimaszewski and Andruszkiewicz (2026) demonstrate how simply rewriting a document in Wikipedia-like style can bypass the FineWeb-Edu filter (Lozhkov et al., 2024), allowing non-educational or otherwise undesirable content to pollute the dataset. This study indicates that more focus on understanding what document features are prioritized by CQFs are needed.

While classifier-based filtering might unintentionally distort the distribution of data, several approaches to intentionally alter the composition of pre-training corpora exist. Source-domain or knowledge-domain re-weighting is commonly used to improve LLM downstream-task performance. Gao et al. (2020) uses heuristically determined data sampling weights to compile multiple source-domains (e.g. GitHub[1] or PubMed[2]) into the PILE dataset. Xie et al. (2023) demonstrates the Domain Reweighting with Minimax Optimization (DoReMi) approach. DoReMi automatically optimizes sampling proportions of source domains with small proxy models, producing a data mixture that improves LLM downstream performance.

Longpre et al. (2024) notes that domain heterogeneity, the diversity of syntactic, topic and reasoning structures, is beneficial for downstream performance, hence source-domain diversity should be prioritized. Additionally, Longpre et al. (2024) shows that the effectiveness of one-dimensional (single quality) filtering depends on the characteristics of the underlying pre-training corpus, indicating that no universally optimal strategy exists.

While data filtering has been extensively studied in the terms of downstream performance, comparatively little attention has been paid to how filtering alters the linguistic variations of pre-training corpora. In a recent study, Wettig et al. (2025) present the WebOrganizer tool to study domain and format topics, indicating that topic re-weighting can improve downstream evaluation tasks as well.

[1] https://github.com/
[2] https://pubmed.ncbi.nlm.nih.gov/

Recent work on the Olmo3 model (Olmo et al., 2026) utilized WebOrganizer, and featured studies on the distribution of topics in pre-training corpora. However, the topic ratios used for constructing the pre-training mixture were optimized purely for downstream performance.

Similarly Myntti et al. (2025) studied how registers and different register compositions influence downstream performance for English language. They extracted datasets consisting of single registers, and used them to train series of small language models, which were subsequently evaluated against benchmarks. Their studies indicate that individual registers as well as combination of registers can lead to changes in model downstream performance. However, the study did not compare the register-based filtering with other CQFs such as FineWeb or FinePDFs-Edu.

In this paper we address the challenges by investigating how educational filters reshape the distribution of web registers and thematic content in pre-training corpus. We compare CQF FinePDFs-Edu with LLM-based Propella-1 annotations, and with the MultiCORE framework.

3 Methods

3.1 Data

Our study focused on the Finnish subset of the "clean" partition of the HPLT 4.0 dataset [3]. The "clean" partition contains documents that have been filtered to remove low-quality and noisy content such as documents consisting mostly of HTML navigation elements, or boilerplate text. HPLT 4.0 is an updated version of HPLT 3.0 described by Oepen et al. (2026), and contains documents in 195 languages, sourced from Internet Archive, the Common Crawl and ArchiveBot. The Finnish subset used in our study consists of 51 million documents. Data was annotated by dataset creators with the MultiCORE register classifier (Henriksson et al., 2024), Propella-1-4b (Idahl et al., 2026), FinePDFs-Edu (Kydlíček et al., 2025) and several other classifiers.

3.2 Data labeling

We used the web register probabilities included in HPLT 4.0 to classify documents with registers. The MultiCORE classifier follows the register scheme developed by Laippala et al. (2023) based on the original work by Biber and Egbert (2018). The scheme is hierarchical with 9 main registers and 47 sub-register categories. Following the approach of Myntti et al. (2025), we use a threshold of 0.4 to assign register labels. A document may receive one or several register labels depending on its linguistic characteristics. Such multi-label documents are called "hybrid" registers. The main register labels and their descriptions are explained in table 1. The sub-registers are more detailed categories; for example, the NA register includes News articles and Personal blogs, the IN register includes Encyclopedia and Research articles, and the OP register includes Reviews and Advice.

Register label	Register description
HI	How-to-instructions
ID	Interactive discussion
IN	Informational description
IP	Informational persuasion
LY	Lyrical
MT	Machine translation
NA	Narrative
OP	Opinion
SP	Spoken

Table 1: Main register labels used in our study

In addition to registers, we study LLM-based Propella-1 annotations `educational_value` and `business_sector`, as well as FinePDFs-Edu scores.

The `educational_value` feature estimates the instructional value of a document, ranging from content with no educational value to structured material that explicitly supports learning through clear explanations, examples, and progressive knowledge building. Propella-1 `educational_value` assigns five categories to documents: `None`, `Minimal`, `Basic`, `Moderate`, `High`.

The Propella-1 `business_sector` is a multi-label feature with 37 possible classes such as "academic research" and "technology software". In the HPLT 4.0 cleaned Finnish subset, the average number of labels per document is 1.9. To simplify our analysis we logged each class in a document as a separate entry, resulting in 98 million business sector entries.

FinePDFs-Edu is a CQF and it assigns a continuous numerical score ranging from 0 to 5, reflecting its suitability as educational content for primary to grade school levels. Documents containing detailed explanation, coherent reasoning, and clear writing receive higher scores, while documents characterized by non-academic writing, poor coherence or superficial treatment of the topic re-

[3] https://hplt-project.org/datasets/v4.0

ceive lower scores.

Propella-1 `educational_value` and FinePDFs-Edu scores were used as present in HPLT 4.0 and were not post processed by us.

3.3 Training and evaluation of small language models

We trained three small language models with the Llama architecture (Touvron et al., 2023), each with 0.6 billion trainable parameters and a training budget of 100 billion tokens. Training was conducted using the NVIDIA Megatron-LM framework (Shoeybi et al., 2019). The models had 22 hidden layers, a hidden size of 1024, and 16 attention heads. We used a sequence length of 4096, a global batch size of 1024, and the Adam optimizer. Hyperparameters were selected based on our initial experiments. All computing was carried out on the CSC-LUMI supercomputer. Additional details are given in Appendix A table 5

To measure the effect of filtering, we trained a model on the unfiltered Finnish subset of HPLT 4.0, which we use as the baseline for downstream performance comparisons. We used Propella-1 to construct a training corpus consisting only of the `high`, `moderate`, and `basic` educational classes, resulting in a collection of 33 billion tokens (BT). We then selected the 33 BT with the highest FinePDFs-Edu scores to construct a corpus of the same size. The two filtered datasets were subsequently used to train separate small language models. All models were trained on 100 BT budget with the same number of steps. However because filtered datasets were smaller than original Finnish subset of HPLT 4.0, the baseline model underwent 1 epoch, and filtered models 3 epochs of training.

To evaluate the three trained models, we used the `LM-evaluation-harness` (Gao et al., 2024) framework. We selected the following tasks:

- ARC:Challenge (Clark et al., 2018): multiple-choice exam questions requiring multi-hop reasoning. We used the `LumiOpen/arc_challenge_mt` implementation [4].
- Belebele (Bandarkar et al., 2024): Multiple-choice reading comprehension task with 900 machine translated and human annotated questions. [5]
- MultiBLiMP (Jumelet et al., 2025): Linguistic minimal pairs test, evaluating subject-verb agreement for number, person, and gender. [6]
- HellaSwag (Zellers et al., 2019): Common sense reasoning task for sentence completion. We used the `LumiOpen/opengpt-x_hellaswagx` implementation [7] (Thellmann et al., 2024)

This selection allowed us to observe different dimensions of model performance. All tasks were run with a zero-shot approach.

4 Results and discussion

We extracted the metadata from the Finnish subset of HPLT 4.0, and divided it into five separate groups respective to the Propella-1 educational_value classes. We compared the register and business_sector distributions between the groups as well as with the distribution of the entire Finnish HPLT 4.0. Table 2 shows the proportion of documents belonging to each Propella-1 class in the dataset.

The recommended use of the FinePDFs-Edu filter is to utilize only the top 10% of the highest scored documents. However, to enable a direct comparison between FinePDFs-Edu and Propella-1, we calculated the FinePDFs-Edu score threshold values which retain the same amount of documents as Propella-1 `educational_value classes`, and repeated the metadata extraction. The thresholds and corresponding Propella-1 classes are: <0.3771 (none), 0.3771-0.8929 (minimal), 0.8929-1.1998 (basic), 1.1998-2.6491 (moderate), >2.6491 (high). Matching the amount of retained data is important from LLM pre-training perspective, where models trained with a fixed token budget can otherwise be exposed to different numbers of epochs depending on the size of the filtered dataset. Controlling the amount of training data therefore allows for direct comparison.

4.1 Comparison of Propella-1 educational values and FinePDFs-Edu

The relationship between FinePDFs-Edu and Propella-1 `educational_value` can be seen in Figure 1. The visible mismatch between both filters

[4] https://huggingface.co/datasets/LumiOpen/arc_challenge_mt

[5] https://huggingface.co/datasets/facebook/belebele

[6] https://huggingface.co/datasets/jumelet/multiblimp

[7] https://huggingface.co/datasets/LumiOpen/opengpt-x_hellaswagx

Educational value	Percent of dataset
None	38.19%
Minimal	34.13%
Basic	17.54%
Moderate	9.27%
High	0.87%

Table 2: Proportions of Propella-1 educational values in the dataset

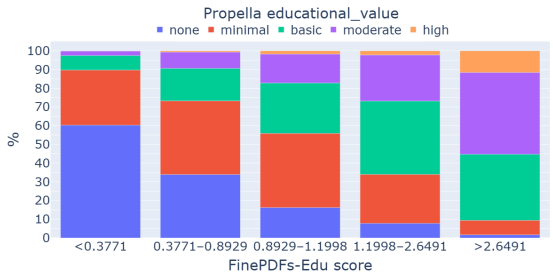


Figure 1: Distribution of propella-1 educational value within FinePDFs-Edu score buckets of matching size

can be seen. At the same time, a trend is observed: with increasing FinePDFs-Edu score, the amount of documents labeled with Propella-1 none and minimal educational classes decreases, while the amount of moderate and high documents increases. It is worth to note that both classifiers mismatch on all educational classes: Propella-1 educational labels none and high can be found throughout all the FinePDFs-Edu score ranges. We further calculated the Spearman rank correlation between both filters by assigning scores (0-4) to Propella-1 educational categories (with 0 = none and 4 for high). The filters correlate with a score of 0.49, which indicates moderate correlation. All these values support finding of weark-to-moderate correlation between filter scores. In addition we calculated Kendalls $\tau_B = 0.37$ and quadratic Cohen's $\kappa = 0.46$ by assigning FinePDFs-Edu to 5 categories as shown in table 3b. Our finding indicates that use of only a single filter might be insufficient to identify the educational content in the corpus. Moreover, the disagreement between filters suggests that both CQF and LLM-based educational filters may misclassify the educational value of documents, either by overlooking educational content or by retaining documents with limited educational value.

4.2 Changes in register distributions

To quantify the effect of educational filtering, we calculated the χ^2 distance between the distributions of registers in full corpus and each educational category. Let p_j denote the overall proportion of observations relating to register *j*:

$$p_j = \frac{\sum_c n_{jc}}{\sum_j \sum_c n_{jc}} \tag{1}$$

where n_{jc} is the number of observations beloning to register *j* and educational category *c*. The proportion of observations belonging to register *j* within educational category *c* is defined as:

$$p_{j|c} = \frac{n_{jc}}{\sum_j n_{jc}}. \tag{2}$$

And finally, the χ^2 distance for educational category *c* is then calculated as:

$$D_{\chi^2}(c) = \sum_j \frac{\left(p_{j|c} - p_j\right)^2}{p_j}, \tag{3}$$

The distance is not normalized to a fixed range. A value zero indicates that the sector distribution within the category is identical to the overall distribution, while larger values indicate greater divergence. As shown in Table 3, the two filters produce broadly similar patterns: register distributions remain close to the full corpus for the lower educational categories, while divergence increases for the highest categories. This can be associated with lower educational classes having more documents, and only the highest classes/scores being very selective. For Propella-1, both moderate and high show strong divergence, whereas for FinePDFs-Edu, it is observed only in the highest score bucket. Interestingly, for both filters, the second category exhibits lower divergence than the first, despite the lowest category containing the most documents.

We evaluated this finding further by inspecting register distributions within the buckets as seen in Figure 2. For simplicity, we show only the 9 main register categories, and due to a large number of hybrid register classes, all hybrids are grouped together. A numerical breakdown of the registers is given in Appendix B tables 6 and 7.

Both filters produce similar register distributions. Figure 2 shows that, for both filters, the proportion of Informational Description (IN) and How-to (HI) documents increases with educational value, while Narrative (NA) and Informational Persuasion (IP) documents show the opposite trend. Both effects are stronger for the Propella-1 filter. In addition, the Interactive Discussion (ID) register is less present in the two highest educational categories. Lyrical (LY) constitutes a marginal part

(a) Educational value	χ^2 distance
None	0.1097
Minimal	0.0467
Basic	0.1226
Moderate	1.0636
High	2.5612

(b) FinePDFs-Edu score	χ^2 distance
<0.3771	0.117457
0.3771–0.8929	0.044733
0.8929–1.1998	0.078526
1.1998–2.6491	0.225581
>2.6491	1.125385

Table 3: χ^2 distances of main registers distribution for a) Propella-1 educational value categories, b) FinePDFs-Edu score buckets

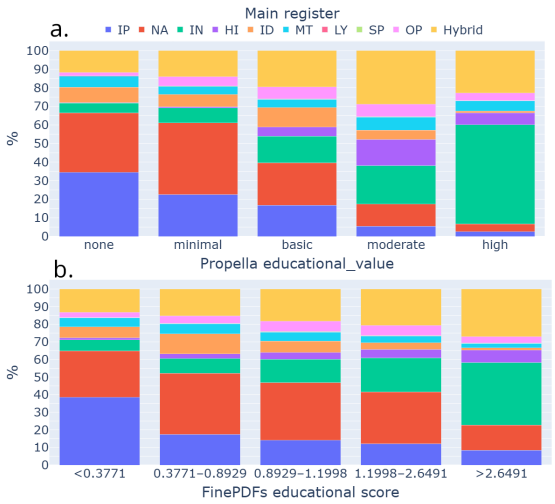


Figure 2: Register distributions for a) Propella-1 educational categories and, b) FinePDFs-Edu score categories of matching size

of the set (<1%), however, LY is mostly present in the lowest educational categories. The changes of distributions of Opinion (OP), Spoken (SP) and Machine Translation (MT) registers are smaller, and their proportion in the educational categories stays relatively constant.

These findings do not completely agree with the results described by Myntti et al. (2025). While numerous studies, such as (Lozhkov et al., 2024), show that educational filters improve the downstream performance of LLMs, Myntti et al. (2025) used register-based filtering. Their results show that datasets consisting only of HI, IN, or OP documents achieve higher downstream performance than datasets consisting of IP or MT documents. However, the presence of MT in the high educational classes and OP in the low educational classes stands in contrast to these findings. Moreover, our finding of an inverse relationship between educational class and the NA register, together with the prevalence of OP in lower educational classes, stands in contrast to Myntti et al. (2025)'s finding that a combination of HI, IN, OP, and NA registers improves the performance of the baseline model. This could indicate that the main source of improved downstream performance is related to the HI and IN registers.

The failure for educational filters to remove MT documents from high educational categories can be explained by filters often prioritizing structured material, explanations, and examples (i.e. the content of the document) rather than style or language quality. The MT register is assigned to a document based on its linguistic characteristics (i.e., translation artifacts). This suggests that educational filters may be prone to overlooking linguistic correctness. To further support this claim, we investigated the distribution of MT hybrid registers. Figure 3 presents the distribution of MT hybrid registers for a) Propella-1 and b) FinePDFs-Edu filters. The label 'multi' refers to hybrid registers combining machine translation with two or more other registers (e.g. MT-HI-IN). The numeric breakdown of prevalence of machine-translated documents can be found in Appendix B, Tables 8 and 9.



Figure 3: Hybrid MT document distribution within a) Propella-1 educational categories and, b) FinePDFs-Edu score categories of matching size

The distribution of MT hybrid registers reveals the same trend as observed within main registers for IN and HI: the amount of MT-IN and MT-HI register documents increases with increasing educational_value.

To further strengthen our finding, we inspected a small random sample of texts with MT-IN or MT-HI register labels, Propella-1 educational value

(a) MT-HI register

Original text: **Käytä kahta kaarevaa viivaa, jotka on yhdistetty typeriin pisteeseen, jotta dorsal fin on kalan takaosassa.**

Translation: "Use two curving lines that are connected to stupid point so the dorsal fin is at the back end of the fish"

(b) MT-IN register

Original text: **Näiden naisten, monien aatelisperheen, varallisuus auttoi häntä perustamaan löytöeläimen ja muita sairaaloita.**

Translation: "The wealth of these women, many noble family's, helped him found a rescue animal and other hospitals."

Figure 4: Examples of documents labeled as (a) MT-HI register with `educational_value`: high, FinePDFs-Edu score: 2.7322; and (b) MT-IN register with `educational_value`: moderate, FinePDFs-Edu score 3.0498.

of `high` or `moderate` and a FinePDFs-Edu score of >2.6491. Many of the examined texts exhibited clear characteristics of machine translation errors. In some of the examples, translation artifacts made the text difficult to interpret by native speakers. Common features included grammatical errors, contextually mismatched words, and left-over non-Finnish words. In some texts individual sentences are fairly coherent, but do not form one cohesive text. The examples of MT register documents are shown in Figure 4.

The example in Figure 4a contains detailed instructions for drawing fish meant for children. This example sentence contains several features that highlight the way educational filters focus on content, not linguistic features. The most notable is the use of the English "*dorsal fin*" rather than the Finnish "*selkäevä*". The noun phrase "*typeriin pisteeseen*" contains two additional errors. The adjective "*typerä*" ("stupid") is clearly a mistranslation. There is also a grammatical error, where the adjective is in its plural form rather than in the singular to match the singular of the noun "*pisteeseen*" ("to a point").

The MT-IN example shown in Figure 4b is describing the life of a historical saint. The sentence contains a number-agreement error ("*monien aatelisperheen*" instead of "*monien aatelisperheiden*") and a likely lexical translation error, where "*löytöeläimen*" ("rescue animal") is semantically incompatible with "*perustaa*" ("to establish").

The above findings suggest that educational filters can be bypassed by MT documents even if only the highest educational categories are used. This in turn can affect linguistic correctness of text generated by LLMs.

4.3 Changes in business sector distributions

Similarly to the analysis of registers presented above, we calculated χ^2 distances for `business_sector` with equations 1, 2, 3 where j now denotes business_sector category. The scores for Propella-1 and FinePDFs-Edu are presented in Table 4. We observe a similar trend to that found in the register analysis: the Propella-1 `high` educational category deviates the most from the distribution of the general Finnish subset. Furthermore, as with the register distributions, the Propella-1 filter results in a stronger deviation from the general distribution than FinePDFs-Edu.

(a) Educational value	χ^2 distance
None	0.1931
Minimal	0.0426
Basic	0.0965
Moderate	0.3785
High	1.9368

(b) FinePDFs-Edu score	χ^2 distance
<0.3771	0.2545
0.3771–0.8929	0.0447
0.8929–1.1998	0.1771
1.1998–2.6491	0.3544
>2.6491	1.3690

Table 4: χ^2 distances of business sector distributions for (a) Propella-1 educational value categories and (b) FinePDFs-Edu score buckets.

Unlike `educational_value`, and similar to register labels, each document can have a single or multiple `business_sector` categories. Figure 5a shows the `business_sector` distribution for the Finnish HPLT 4 dataset. We treat each label as separate entry i.e., we split each multilabel-document pair to label-document pairs. The most populated categories within the Finnish subset of HPLT 4.0 are: general interest, media and entertainment, and government public. The distribution changes significantly for both filters. Figure 5a shows the distribution for the Propella-1 educational category `high` and 5c the distribution for the respective FinePDFs-Edu score range (the same proportion of the text as educational category `high`). Both Propella-1 and FinePDFs-Edu increase the proportion of academic

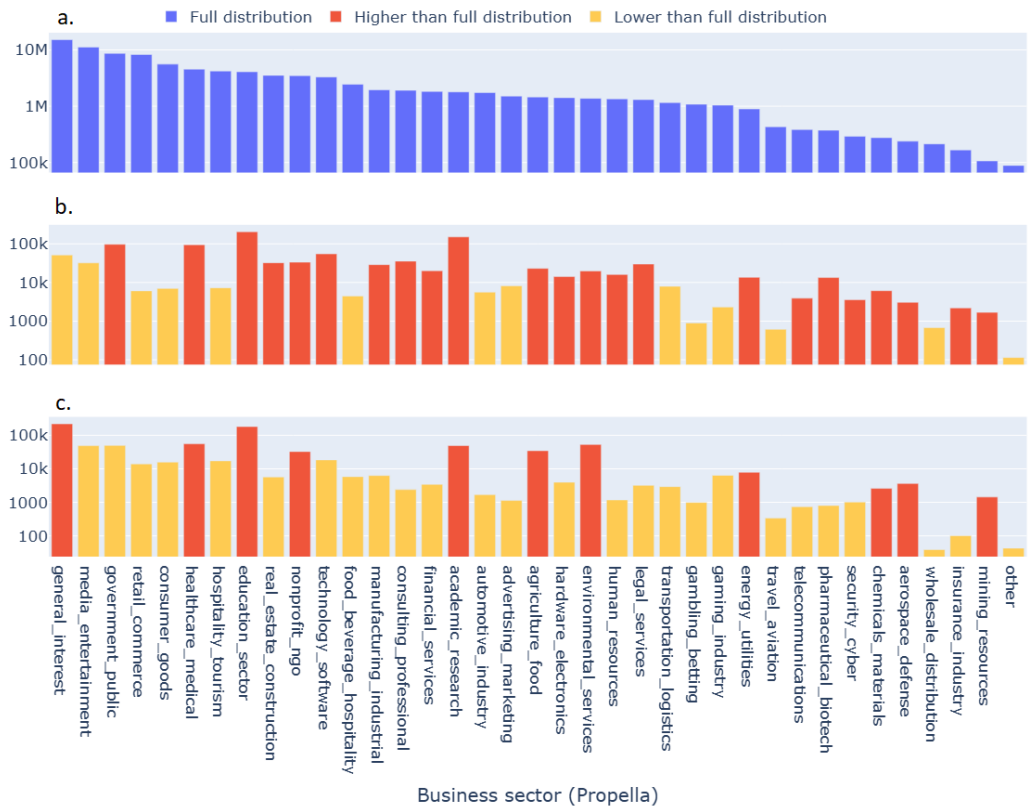


Figure 5: Distributions of business sectors in a) Finnish subset of HPLT 4.0, b) Subset with educational_value: high; c) Subset with FinePDFs score >2.6491. The bar colors on plots b) and c) indicate if the proportion of given sector is higher or lower than for the general distribution.

research documents (8.0x, 6.4x respectively), educational sector (4.8x, 2.9x respectively), as well as 8 others. However, the proportion of "general interest" documents is higher for FinePDFs-Edu than for unfiltered Finnish subset of HPLT 4.0, while for Propella-1 it is lower. In addition, Propella-1 has an increased proportion of 12 business sectors which are less present in the FinePDFs-Edu filtered data.

The distribution of business sectors within educational categories in shown in Appendix C tables 10 and 11. Appendix D Figure 7 shows a version of Figure 5 with the top 10 % of texts sorted by FinePDFs educational score.

These findings contrast with the register analysis, where both filters showed largely consistent shifts towards the same types of registers in highly educational documents. This suggests that, while the two filters identify similar document characteristics associated with educational value such as step-by-step instructions and explanations of concepts, they differ more substantially in how they assess the thematic content of documents. Such differences in thematic composition may in turn affect the knowledge represented in the resulting models and consequently, model performance across different types of evaluation tasks.

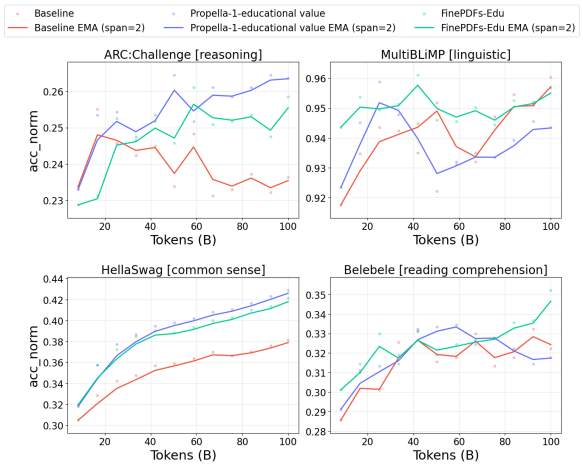


Figure 6: Downstream performance of three trained language models. The dots represent checkpoint performance, while lines show the exponential moving average. The standard error for ARC:Challenge, MultiBLiMP, HellaSwag and Belebele tasks are 0.012, 0.009, 0.004 and 0.015 respectively.

4.4 Downstream performance evaluation

We evaluated three small language models: a baseline model which was trained on the unfiltered Finnish subset of HPLT 4.0, and two models trained on subsets filtered by Propella-1 and FinePDFs-Edu. Figure 6 shows the performances of the evaluation tasks.

The effect of data filtering varies across evaluation tasks. For multi-hop and common sense reasoning tasks (ARC:Challenge and HellaSwag respectively), the model trained with data filtered with Propella-1 outperforms the two others. However, Propella-1 shows drop in performance for the MultiBLiMP linguistic minimal pairs task, which does not recover till end of training. Additionally, for the reading comprehension task (Belebele), all three models perform similary during training, except for last three checkpoints, where FinePDFs-Edu model outperforms two other. The improved performance of filtered models on multi-hop and common sense reasoning tasks can be associated with increased amount of instructional and informative documents in training corpora.

5 Conclusions

We analyzed how educational filters affect the distribution of linguistic variation and thematic content (business sectors) in a pre-training corpus using Propella-1 annotations, FinePDFs-Edu, the MultiCORE register (genre) framework, and the Finnish subset of HPLT 4.0. Our results show that educational filters shift the distribution towards instructional and informative registers, and away from narrative and persuasive documents. Similarly, filtering increases the representation of academic, technical and medical business sectors.

Our analysis of machine-translated documents further shows that high educational scores do not always correspond to linguistically well-formatted text. Documents containing grammatical and lexical errors are present even among the highest educational categories of both filters. We determined that educational filters focus on content characteristics and ignore linguistic quality that introduces risk of including machine-translated documents in the filtered corpus.

Finally, by studying data distribution changes, training small language models, and evaluating them we show that educational filters have both shared and distinct effects on corpus and model performance. These findings highlight the importance of considering linguistic and thematic composition alongside downstream performance when evaluating pre-training data filtering strategies.

Limitations

In our work we analyze the effect of educational filters on the change of distributions of linguistic variance as well as business categories of documents within the HPLT 4.0 corpus. However, our study focuses only on the Finnish subset of the dataset. Additional studies should be carried out to fully describe the effect of filters on broader range of languages. Moreover we used only two educational filters: FinePDFs-Edu and Propella-1. Additional studies using other filters could clarify our claims and findings. For qualitative evaluation of machine translated text we only consider a small sample of 15 randomly selected texts, reviewed by one native speaker. Lastly, we train three small language models, but do not carry out model size ablation studies - observed effects might differ for larger model sizes. We carried out tests with multiple choice formula for ARC:Challenge and HellaSwag tasks, meanwhile a cloze formula test could result in higher values.

Acknowledgments

This project has received funding from OpenEuroLLM and LLMs4EU projects, under Digital Europe Programme grant agreements No 101195233 and 101198470, and was supported by the Research Council of Finland through several projects: FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union– NextGenerationEU instru ment), “Mechanisms of Register Variation in Mas sively Multilingual Web-Scale Corpora” (project 362459), and the “Centre of Excellence Human Diversity through Contacts” (project 374223). We gratefully acknowledge CSC– IT Center for Science Ltd. for providing computational resources. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori

Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Preprint*, arXiv:2402.16827.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1):9–37.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2024. Automatic register identification for the open web using multilingual deep learning. *Preprint*, arXiv:2406.19892.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pásztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan Garcia Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 83 others. 2026. Apertus: Democratizing open and compliant LLMs for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46877–46955, San Diego, California, United States. Association for Computational Linguistics.

Maximilian Idahl, Benedikt Droste, Björn Plüster, and Jan Philipp Harries. 2026. propella-1: Multi-property document annotation for llm data curation at scale. *Preprint*, arXiv:2602.12414.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Mateusz Klimaszewski and Piotr Andruszkiewicz. 2026. Is a document educational or just wikipedia-style? – pitfalls of classifier-based quality filtering. *Preprint*, arXiv:2605.23721.

Hynek Kydlíček, Guilherme Penedo, and Leandro von Werra. 2025. Finepdfs. `https://huggingface.co/datasets/HuggingFaceFW/finepdfs_edu`.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079. Funding Information: Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital. Funding was provided by Academy of Finland (Grant No. 331297), Emil Aaltosen säätiö, National Science Foundation (Grant No. 1147581). Publisher Copyright: © 2022, The Author(s).

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2025. Datacomp-lm: In search of the next generation of training sets for language models. *Preprint*, arXiv:2406.11794.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. Register always matters: Analysis of LLM pretraining data through the lens of language variation. In *Second Conference on Language Modeling*.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 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. *Preprint*, arXiv:2511.01066.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards cross-lingual llm evaluation for european languages. *Preprint*, arXiv:2410.08928.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, Xu Han, and Zhiyuan Liu. 2025. Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data. *Preprint*, arXiv:2505.05427.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pretraining data curation. *Preprint*, arXiv:2502.10341.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pages 69798–69818. Curran Associates, Inc.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

A Appendix

The small language models were trained with the following hyperparameters:

Parameter	Value
Number of layers	28
Hidden size	1024
Number of attention heads	16
Sequence length	4096
Learning rate	0.002
Learning rate decay type	WSD
Decay style	linear
Warmup iterations	2000
Cooldown fraction	0.2
Weight decay	0.1
Optimizer	Adam
Adam β_1	0.9
Adam β_2	0.95
Adam ϵ	1e-8
Global batch size	512
Micro batch size	2
Training token budget	100BT
Tokenizer	openeurollm/tokenizer- 256k

Table 5: Most important hyperparameter of trained models. The openerollm tokenizer can be found under https://huggingface.co/openeurollm/tokenizer-256k

B Appendix

Table 6 shows the register composition of each Propella-1 educational category. The same composition for FinePDFs-Edu score categories is shown in table 7.

	none	minimal	basic	moderate	high
IP	34.42 %	22.57 %	16.77 %	5.51 %	2.64 %
NA	32.04 %	38.58 %	22.92 %	11.95 %	4.09 %
IN	5.28 %	8.05 %	14.26 %	20.69 %	53.40 %
HI	0.23 %	0.59 %	4.91 %	14.08 %	6.40 %
ID	8.31 %	6.67 %	10.60 %	4.98 %	0.95 %
MT	5.99 %	4.22 %	4.30 %	7.05 %	5.54 %
LY	0.16 %	0.03 %	0.01 %	0.02 %	0.02 %
SP	0.09 %	0.20 %	0.29 %	0.28 %	0.18 %
OP	1.74 %	5.03 %	6.38 %	6.63 %	3.95 %
Hybrid	11.73 %	14.07 %	19.57 %	28.81 %	22.83 %

Table 6: Proportion of registers within each Propella-1 educational category

	<0.3771	0.3771–0.8929	0.8929–1.1998	1.1998–2.6491	>2.6491
IP	38.45 %	17.44 %	14.17 %	12.20 %	8.43 %
NA	26.49 %	34.77 %	32.72 %	29.39 %	14.24 %
IN	6.21 %	8.33 %	13.29 %	19.34 %	35.55 %
HI	1.23 %	2.59 %	3.73 %	4.76 %	7.13 %
ID	6.23 %	11.38 %	6.52 %	3.85 %	1.23 %
MT	5.13 %	5.77 %	5.07 %	3.74 %	2.57 %
LY	0.04 %	0.10 %	0.11 %	0.07 %	0,01 %
SP	0.06 %	0.20 %	0.33 %	0.29 %	0.20 %
OP	2.91 %	4.25 %	5.85 %	5.68 %	3.57 %
Hybrid	13.25 %	15.16 %	18.21 %	20.69 %	27.06 %

Table 7: Proportion of registers within each FinePDFs-Edu score category

Tables 8 and 9 show what percentage of texts in each educational score category belong to the MT register or a hybrid register containing MT. The label 'multi' denotes hybrid registers composed of three or more main registers.

	none	minimal	basic	moderate	high
MT	5.9933 %	4.2243 %	4.2998 %	7.0457 %	5.5436 %
IP-MT	2.2730 %	0.5112 %	0.2784 %	0.1285 %	0.1038 %
IN-MT	0.1015 %	0.2167 %	0.3308 %	0.8144 %	1.9455 %
HI-MT	0.0122 %	0.0839 %	0.2221 %	1.3349 %	0.7746 %
MT-OP	0.2636 %	0.1226 %	0.1020 %	0.1345 %	0.0217 %
MT-NA	0.1185 %	0.1035 %	0.0340 %	0.0188 %	0.0072 %
multi	0.0782 %	0.0138 %	0.0152 %	0.0451 %	0.0408 %
ID-MT	0.0325 %	0.0329 %	0.0365 %	0.0197 %	0.0256 %
LY-MT	0.0082 %	0.0016 %	0.0003 %	0.0000 %	0.0000 %
MT-SP	0.0003 %	0.0010 %	0.0010 %	0.0016 %	0.0009 %

Table 8: Amount of machine translated registers as percentage of texts in Propella educational categories

	<0.3771	0.3771–0.8929	0.8929–1.1998	1.1998–2.6491	>2.6491
MT	5.1300 %	5.7697 %	5.0669 %	3.7396 %	2.5721 %
IP-MT	2.1767 %	0.6489 %	0.2376 %	0.1019 %	0.0248 %
IN-MT	0.0519 %	0.2747 %	0.5814 %	0.4845 %	0.3064 %
HI-MT	0.1037 %	0.2697 %	0.2435 %	0.2878 %	0.1924 %
MT-OP	0.3588 %	0.0811 %	0.0378 %	0.0176 %	0.0074 %
MT-NA	0.0921 %	0.1179 %	0.0613 %	0.0233 %	0.0027 %
multi	0.0824 %	0.0200 %	0.0132 %	0.0113 %	0.0134 %
ID-MT	0.0243 %	0.0452 %	0.0324 %	0.0179 %	0.0029 %
LY-MT	0.0011 %	0.0064 %	0.0060 %	0.0009 %	0.0002 %
MT-SP	0.0008 %	0.0010 %	0.0008 %	0.0002 %	0.0000 %

Table 9: Amount of machine translated registers as percentage of texts in FinePDFs educational score categories

C Appendix

Figure 7 shows how the proportion of business sector labels is affected by data selection. Figure 7a shows the baseline distribution of the whole dataset, while 7b shows distribution in documents labelled `high` by the Propella-1 educational score classifier. 7c shows the distribution in the 10 % of texts with the highest FinePDFs-Edu score. Bar colors in plots b and c indicate whether the proportion of the given sector is higher or lower than in the whole dataset.

D Appendix

Table 10 contains a breakdown of Propella-1 business sectors within the 5 educational categories given by Propella-1. Table 11 shows the same for educational score categories from FinePDFs-Edu.

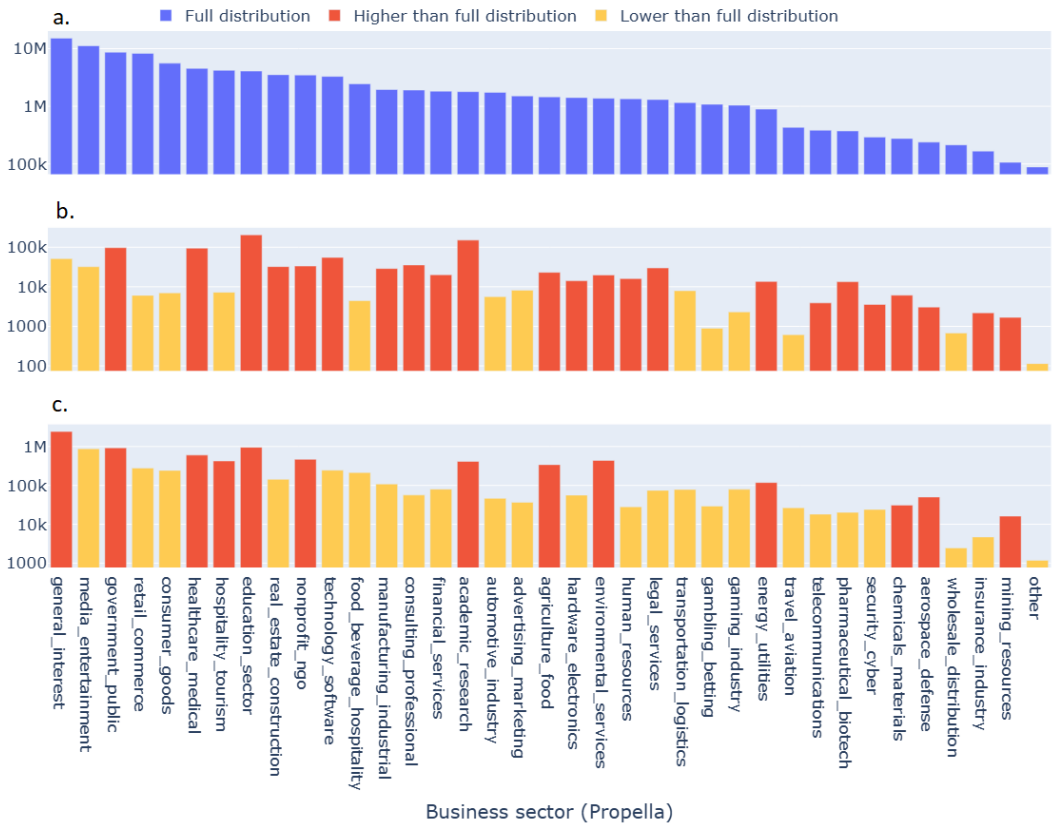


Figure 7: Distributions of business sectors in a) Finnish subset of HPLT 4.0, b) Subset with Propella-1 educational value high, c) Top 10 % of documents with the highest FinePDFs-Edu score

Business sector	none	minimal	basic	moderate	high
general_interest	39.20 %	28.85 %	20.21 %	11.40 %	0.34 %
media_entertainment	47.25 %	38.72 %	10.44 %	3.30 %	0.29 %
government_public	18.72 %	51.38 %	20.63 %	8.15 %	1.12 %
retail_commerce	55.59 %	30.64 %	11.12 %	2.58 %	0.07 %
consumer_goods	50.14 %	29.85 %	14.71 %	5.17 %	0.13 %
healthcare_medical	15.34 %	35.67 %	27.93 %	18.98 %	2.08 %
hospitality_tourism	48.72 %	33.11 %	14.70 %	3.30 %	0.17 %
education_sector	11.71 %	47.26 %	24.46 %	11.55 %	5.03 %
real_estate_construction	43.10 %	32.69 %	15.38 %	7.92 %	0.92 %
nonprofit_ngo	26.09 %	46.91 %	19.29 %	6.75 %	0.96 %
technology_software	17.36 %	38.28 %	23.55 %	19.15 %	1.67 %
food_beverage_hospitality	27.56 %	25.84 %	25.53 %	20.89 %	0.18 %
manufacturing_industrial	32.14 %	40.74 %	17.12 %	8.52 %	1.48 %
consulting_professional	24.95 %	36.12 %	20.08 %	16.99 %	1.86 %
financial_services	19.43 %	39.53 %	25.02 %	14.92 %	1.10 %
academic_research	3.39 %	30.20 %	29.59 %	28.40 %	8.42 %
automotive_industry	40.82 %	31.92 %	17.46 %	9.48 %	0.32 %
advertising_marketing	44.34 %	28.58 %	15.98 %	10.56 %	0.54 %
agriculture_food	15.71 %	38.29 %	27.50 %	16.92 %	1.59 %
hardware_electronics	25.27 %	45.57 %	17.29 %	10.87 %	1.00 %
environmental_services	9.70 %	41.29 %	33.93 %	13.65 %	1.44 %
human_resources	37.55 %	30.93 %	17.82 %	12.51 %	1.19 %
legal_services	12.30 %	40.96 %	25.40 %	19.06 %	2.29 %
transportation_logistics	29.72 %	45.88 %	17.58 %	6.14 %	0.69 %
gambling_betting	50.26 %	31.05 %	13.54 %	5.06 %	0.08 %
gaming_industry	51.68 %	33.49 %	10.10 %	4.51 %	0.22 %
energy_utilities	17.30 %	44.71 %	25.02 %	11.45 %	1.52 %
travel_aviation	38.56 %	41.00 %	15.91 %	4.40 %	0.14 %
telecommunications	20.07 %	49.42 %	19.94 %	9.56 %	1.02 %
pharmaceutical_biotech	20.79 %	24.50 %	25.26 %	25.87 %	3.58 %
security_cyber	12.56 %	32.98 %	27.72 %	25.51 %	1.23 %
chemicals_materials	14.53 %	38.81 %	27.86 %	16.59 %	2.22 %
aerospace_defense	8.00 %	51.12 %	26.46 %	13.14 %	1.29 %
wholesale_distribution	60.66 %	28.20 %	7.47 %	3.35 %	0.32 %
insurance_industry	16.23 %	33.02 %	30.60 %	18.84 %	1.31 %
mining_resources	19.28 %	46.08 %	21.97 %	11.10 %	1.57 %
other	87.17 %	7.46 %	3.54 %	1.69 %	0.13 %

Table 10: Distribution of Propella-1 business sectors categories within Propella-1 educational value categories

Business sector	<0.3771	0.3771–0.8929	0.8929–1.1998	1.1998–2.6491	>2.6491
general_interest	7.89 %	18.22 %	20.02 %	24.13 %	25.62 %
media_entertainment	11.51 %	12.24 %	10.63 %	9.10 %	5.74 %
government_public	5.28 %	11.03 %	11.33 %	9.58 %	5.83 %
retail_commerce	15.27 %	5.28 %	3.37 %	2.89 %	1.65 %
consumer_goods	10.11 %	3.55 %	2.27 %	2.48 %	1.86 %
healthcare_medical	2.48 %	5.37 %	6.62 %	5.99 %	6.51 %
hospitality_tourism	4.62 %	3.86 %	4.24 %	4.46 %	2.03 %
education_sector	1.37 %	4.18 %	6.57 %	8.46 %	21.24 %
real_estate_construction	5.06 %	3.32 %	2.24 %	1.51 %	0.67 %
nonprofit_ngo	2.20 %	3.75 %	5.14 %	4.79 %	3.84 %
technology_software	4.11 %	3.24 %	2.45 %	2.49 %	2.16 %
food_beverage_hospitality	2.13 %	2.58 %	3.13 %	2.30 %	0.69 %
manufacturing_industrial	2.74 %	1.70 %	1.43 %	1.12 %	0.74 %
consulting_professional	2.25 %	2.29 %	1.40 %	0.60 %	0.28 %
financial_services	1.97 %	2.25 %	1.45 %	0.84 %	0.40 %
academic_research	0.35 %	1.78 %	3.54 %	3.99 %	5.76 %
automotive_industry	3.08 %	1.30 %	0.72 %	0.49 %	0.20 %
advertising_marketing	2.50 %	1.30 %	0.64 %	0.39 %	0.13 %
agriculture_food	0.81 %	1.24 %	2.15 %	3.35 %	4.06 %
hardware_electronics	2.62 %	0.90 %	0.52 %	0.57 %	0.47 %
environmental_services	0.39 %	1.00 %	2.47 %	4.23 %	6.20 %
human_resources	1.85 %	1.46 %	0.82 %	0.30 %	0.14 %
legal_services	1.22 %	1.71 %	1.15 %	0.78 %	0.38 %
transportation_logistics	1.17 %	1.30 %	1.14 %	0.83 %	0.34 %
gambling_betting	2.04 %	0.72 %	0.33 %	0.31 %	0.12 %
gaming_industry	1.60 %	0.81 %	0.55 %	0.79 %	0.75 %
energy_utilities	0.59 %	0.92 %	1.33 %	1.21 %	0.93 %
travel_aviation	0.44 %	0.50 %	0.42 %	0.29 %	0.04 %
telecommunications	0.59 %	0.33 %	0.21 %	0.19 %	0.09 %
pharmaceutical_biotech	0.25 %	0.51 %	0.49 %	0.21 %	0.09 %
security_cyber	0.27 %	0.36 %	0.25 %	0.25 %	0.12 %
chemicals_materials	0.28 %	0.25 %	0.32 %	0.31 %	0.31 %
aerospace_defense	0.09 %	0.27 %	0.34 %	0.51 %	0.43 %
wholesale_distribution	0.43 %	0.13 %	0.05 %	0.03 %	0.00 %
insurance_industry	0.21 %	0.20 %	0.10 %	0.05 %	0.01 %
mining_resources	0.07 %	0.10 %	0.16 %	0.16 %	0.17 %
other	0.17 %	0.06 %	0.03 %	0.01 %	0.01 %

Table 11: Distribution of Propella-1 business sectors categories within FinePDFs-Edu score categories

WiNCC: A Reproducible Pipeline for Constructing a Wikipedia-enriched Nepali Cross-lingual Corpus

Ahrnish P Dahal[1], Pushkar Baranwal[2], Amit Pundir[3], Geetika Jain Saxena[3], Sanjeev Singh[2]
Cluster Innovation Centre, University of Delhi[1]
Institute of Informatics and Communication, University of Delhi[2]
Maharaja Agrasen College, University of Delhi[3]

Abstract

WiNCC is a reproducible and automatic construction pipeline for building English–Nepali cross-lingual retrieval corpus that is difficulty-aware, relying on public Wikipedia and Wikidata.

Part 1 reproduces a CLIRMatrix-style baseline. Nepali Wikipedia was crawled, articles resolved to a Wikidata QID then intersected with corresponding Wikidata sitelinks to obtain Nepali–English query pairs and related English document corpus. These were sampled to gold pairs and padded with random English distractors into a document baseline corpus.

Part 2 adds a data-enrichment stage to Part 1 with multi-query, class-matched hard negatives. We compare the produced corpus against existing cross-lingual retrieval resources, run cross-lingual retrieval baselines with three distinct multilingual embedding models (BGE-M3, multilingual-E5-large, and Qwen3-Embedding-4B), analyze slice-wise quality-metadata and query-type-wise field metrics supporting the corpus's use as a difficulty-aware benchmark.

Through retrieval tasks on our corpus, we also find (1) a 27.8–41.5 point NDCG@10 drop from English to Nepali, (2) descriptions are 5-6x harder than titles, and (3) hard negatives are 2x stickier than random distractors.

1 Introduction

Building evaluation resources for low-resource, cross-lingual retrieval by hand requires bilingual annotators or expensive machine-translation and labelling step. For this work, a retrieval corpus consisting of English Wikipedia documents was crawled; the query language varies (obtained from title queries for English and Nepali). We follow the automatic language-linking methodology of CLIRMatrix [Sun and Duh, 2020] and its sentence-centric predecessor WikiCLIR [Sasaki et al., 2018] to construct an English–Nepali query bucket directly from Wikipedia's per-language editions and Wikidata's cross-lingual entity graph, with no need for translation or human relevance judgment. The corpus construction procedure is fully automatic from public APIs, and reported stage-by-stage, with an aim to make it directly reproducible. Part 2 (also referred to as v2) adds a data-enrichment and quality-annotation layer (§4) to the Part 1 output, aligning with Web-as-Corpus call for work on data enrichment and quality.

Concretely, we contribute:

(1) Reproducible corpus creation pipeline (§3);
(2) Enrichment and hard-negative mining (§4);
(3) Cross-lingual retrieval baselines (§5);
(4) Metadata-sliced evaluation (§6).

Extended per-slice tables and all supporting visualizations are provided in Appendix A.

2 Related Work

This section discusses the relevant literature and previous works referenced and also compares them against our work.

Web corpus construction and filtering. OSCAR [Suárez et al., 2019] worked on a multi-threaded pipeline to clean and classify Common Crawl by language built especially for medium-resource infrastructures; CCNet [Wenzek et al., 2020] extracted massive high-quality monolingual datasets from Common Crawl for a variety of languages; and ParaCrawl [Bañón et al., 2020] mines web-scale parallel corpora by crawling the web.

Automatic cross-lingual dataset construction. CLIRMatrix [Sun and Duh, 2020] used Wikipedia's inter-language structure, via Wikidata [Vrandečić and Krötzsch, 2014] to construct a collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval. WikiCLIR [Sasaki et al., 2018] is the precedent that automatically extracted German-English retrieval dataset for Cross-Language Information Retrieval. WikiMatrix [Schwenk et al., 2021] mines Wikipedia for

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 131–138
October 29, 2026 ©2026 Association for Computational Linguistics

parallel sentences rather than paired queries and documents.

Multilingual retrieval benchmarks. Widely used multilingual retrieval benchmarks MIRACL [Zhang et al., 2023] and Mr. TyDi [Zhang et al., 2021] rely on human relevance judgments and their language set does not include Nepali. Cross-lingual QA datasets such as XOR-QA [Asai et al., 2021], TyDi QA [Clark et al., 2020], and MKQA [Longpre et al., 2021] target human-authored-QA problem. Table 1 compares **WiNCC** against these existing resources.

Multilingual representation and dense retrieval. XLM-R [Conneau et al., 2020] established pretraining multilingual language models at scale improves performance for a wide range of cross-lingual transfer tasks; Sentence-BERT [Reimers and Gurevych, 2019] showed that Siamese fine-tuning turns such encoders into fast, comparable sentence embeddings, and LaBSE [Feng et al., 2022] extended this to a language-agnostic objective. For retrieval, Dense Passage Retrieval [Karpukhin et al., 2020] proved that dual-encoder dense retrieval can outperform sparse lexical methods for open-domain QA, later extended multilingually. BGE-M3 [Chen et al., 2024], multilingual-E5-large [Wang et al., 2024], and Qwen3-Embedding [Zhang et al., 2025] represent the current generation of broad multilingual embedding models built on this lineage; we use all three only as external instruments for retrieval baselines.

Dense retrieval and hard-negative mining. A separate line of work improves dense retrievers: ANCE [Xiong et al., 2021] applies nearest neighbour negative contrastive learning for dense retrieval and RocketQA [Qu et al., 2021] works on cross-batch negatives; ColBERT [Khattab and Zaharia, 2020] implements late interaction in the scoring stage to make finer distinctions between close candidates; and Contriever [Izacard et al., 2022] mines hard negatives entirely unsupervised via contrastive pretraining.

3 Part 1: Baseline Corpus Construction

This section discusses the construction process and stage-wise counts of the various components in our baseline dataset.

Crawl and Wikidata linking. The pipeline crawls Nepali Wikipedia via MediaWiki endpoints. A full crawl checked **42,434** Nepali articles and resolved **30,689** to a Wikidata QID. Each QID was batch-queried against Wikidata (in batches of 50), yielding **26,431** Nepali–English pairs (∼86% of QID-resolved articles have an English counterpart).

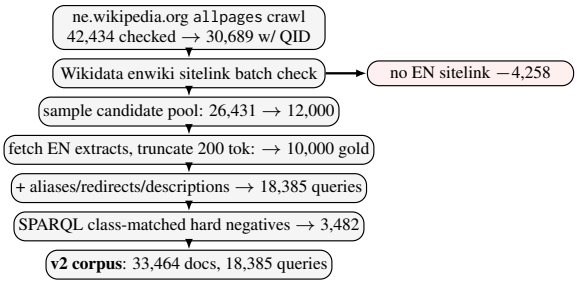


Figure 1: End-to-end pipeline (Part 1 + Part 2), with per-stage counts.

Sampling and gold document collection. From the 26,431 pairs we sampled 12,000 pairs (`random_state=42`) as a safe pool (in case of rate limit failures during extraction) against a set target of 10,000 gold pairs. English extracts were fetched in batches of 20, truncated to 200 whitespace tokens, following CLIRMatrix [Sun and Duh, 2020], till 10,000 gold documents were collected. (Min/Mean/Max = 4 / 114.9 / 200 tokens).

Baseline distractor construction. Distractors were drawn via `list=random`, deduplicated against gold titles and required to have ≥20 tokens, truncated identically, until exactly 20,000 were collected (Min/Mean/Max = 20 / 77.0 / 200 tokens). The baseline corpus contains **30,000** documents (10,000 gold + 20,000 distractors).

4 Part 2: Multi-Query, Hard Negatives, and Metadata

This section discusses the procedure and numbers associated with the Part 2 dataset. Part 2 expands the dataset across three key dimensions:

Expanded Multi-Query Retrieval. To reflect realistic search behaviours, three additional Nepali queries were mapped to gold entities via `wbgetentities`: Wikidata aliases (1,035), Wikipedia redirects (4,277), and Wikidata descriptions (3,073). Combined with 10,000 original titles, this yields **18,385 query formulations** across 10,000 gold documents.

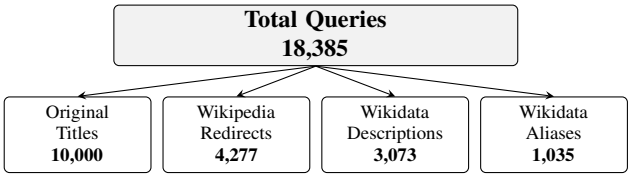


Figure 2: Distribution of the 18,385 total queries across their source categories.

Class-Based Hard Negatives. Topically close

Corpus	#Langs	Nepali	Construction	Relevance	Hard neg.	Multi-query	Metadata
WikiCLIR [Sasaki et al., 2018]	1 pair	no	Wikipedia-derived	binary/graded, auto	no	no (first-sentence query)	no
CLIRMatrix [Sun and Duh, 2020]	∼139	incl., unaudited	Sitelinks	binary, auto	no	no (title only)	no
MIRACL [Zhang et al., 2023]	18	no	Human judgments	graded	implicit (pooled)	no	no
Mr. TyDi [Zhang et al., 2021]	11	no	Human judgments	binary	implicit (pooled)	no	no
WikiMatrix [Schwenk et al., 2021]	1620 pairs	sentence-level	Margin mining	n/a (bitext)	n/a	n/a	no
XOR-QA [Asai et al., 2021]	7	no	Human-authored	binary	no	no	no
WiNCC (ours)	1 pair	**dedicated**	Wikidata/WP	binary, auto	**SPARQL class-matched**	**1.84 qry/entity**	**5 fields**

Table 1: WiNCC vs. existing cross-lingual retrieval and bitext resources. "Multi-query" = more than one query form per gold document; "Metadata" = attached per-instance difficulty/quality fields.

distractor documents were mined using SPARQL queries against primary Wikidata classes. Filtering for documents with $\geq$ 20 tokens yielded 3,482 **hard negatives** (38.5% yield) from 9,072 initial candidate titles.

Per-Instance Metadata. Each instance is enriched with four metadata fields: `alias_count`, `ambiguity_level`[1], `entity_category`, and `lexical_overlap`.

Part 2 corpus assembly. The final v2 corpus merges gold documents, hard negatives, and baseline random distractors: 10,000 gold, 3,482 hard negatives, 19,982 random distractors (18 documents fell through during rate limiting), for a total of **51,849** entities : **33,464 documents** against **18,385 query formulations** (Table 2).

Component	Part 1	Part 2 (v2)
Gold documents	10,000	10,000
Query formulations	10,000	18,385
Hard negatives	–	3,482
Random distractors	20,000	19,982
Total corpus size	**30,000**	**51,849**

Table 2: Component counts, baseline vs. v2 corpora.

Dataset Access and Structure

The complete WiNCC corpus, including both baseline and enriched variants along with pipeline scripts are available at WiNCC.

The data release is organized into two directories, corresponding to the construction stages described in §3 and §4

- `Part 1/`: Contains the baseline 30,000-document corpus.
  - `query_pairs_final.csv`: The 10,000 Nepali–English title query pairs.
  - `gold_docs_truncated.json`: The 10,000 gold English Wikipedia extracts, truncated to 200 whitespace tokens.
  - `distractor_docs.json`: The 20,000 random English distractor documents.
  - `corpus_full.csv`: The complete baseline corpus, merging gold and distractor documents with their binary relevance labels.
- `Part 2/`: Contains the enriched v2 corpus with multi-query formulations and hard negatives.
  - `multi_query_benchmark_enriched.csv`: The full 18,385 query formulations (titles, redirects, aliases, and descriptions) with attached per-instance metadata (ambiguity, category, lexical overlap, etc.).
  - `hard_negative_titles_by_gold.json`: Mapping of the 3,482 class-matched hard-negative document titles to their corresponding gold entities.
  - `corpus_full_v2.csv`: The final 33,464-document corpus, combining gold documents, hard negatives, and random distractors with their respective `doc_type` labels.
  - `wikidata_entity_details.json`: Cached Wikidata entity metadata (aliases, descriptions, and QID properties) used during the enrichment stage.

This results in a total of 51,849 entities. (inclusive of query formulations). The repository also includes `wincc_creation.ipynb`, a notebook for the complete pipeline of corpus creation. All files are provided in plain CSV and JSON formats for ease of use with standard retrieval toolkits.

5 Cross-Lingual Retrieval Baselines

All retrieval experiments use the same English document corpus constructed in §3 and §4 . We vary only the query language, comparing retrieval task with English query formulations against Nepali query formulations. We report retrieval baselines from three architecturally distinct multilingual embedding models – BGE-M3 [Chen et al., 2024], multilingual-E5-large [Wang et al., 2024], and Qwen3-Embedding-4B [Zhang et al., 2025] – encoding all v2-corpus documents and both English and Nepali queries, retrieving (mono- and cross-lingual) with FAISS index [Johnson et al., 2019].

[1]Ambiguity is overwhelmingly low: 18,336/18,385 query rows (99.7 %) fall in the low bucket, with only 43 medium and 6 high rows.

mE5 uses its `"query:"/"passage:"` prefix convention; Qwen3-Embedding uses its query-instruction prompt; BGE-M3 uses symmetric encoding.

Model	Query	R@1	R@10	R@100	NDCG@10
BGE-M3	EN	0.945	0.990	0.995	0.971
	NE	0.502	0.706	0.831	0.603
mE5-large	EN	0.950	0.991	0.995	0.973
	NE	0.449	0.671	0.820	0.558
Qwen3-Emb-4B	EN	0.971	0.993	0.996	0.984
	NE	0.627	0.783	0.878	0.706

Table 3: Overall retrieval metrics by query language, three models. All documents in English. EN = English query, NE = Nepali query. Full per-slice breakdowns are in Appendix A.

All three models are near-ceiling on English queries and show a consistent drop on Nepali queries. Qwen3-Embedding-4B is consistently strongest for English and Nepali queries, mE5-large consistently weakest for Nepali queries. This convergence across three independent models confirms both that gold labels are correct and recoverable, and that the difficulty effect is a property of the corpus's query-language design.

6 Metrics Sliced by Query Metadata

Query-metadata and their associated metrics were a revealing study. Queries were divided in the following ways.

By query type. Titles are consistently easiest (NDCG@10 0.66–0.81), redirects and aliases close behind, and descriptions by far hardest (0.16–0.26): as a paraphrase requires semantic search tasks beyond entity lookup, so retrieval degrades. Identical results ordering for all three models highlights a corpus property, not a single model's weakness.

By ambiguity. (Fig. 3c.) Higher-ambiguity queries score better (0.60–0.71 low vs. 0.74–0.89 high). This is not conclusive as the ambiguity bucket was very tiny (n=6) dominated by short, distinct proper names with easy string matches.

By entity category. `Taxon` entities collapse to NDCG@10 0.08–0.12 for every model, roughly one-sixth the average, suspected due to asymmetric pre-training of models on scientific nomenclature. Administrative and human categories sit close to the pooled mean.

By lexical overlap. Devanagari queries vs. Latin titles show uniformly low character overlap (0.04–0.08), yet the some-overlap slice ($\geq$0.05, n=10,830, mostly transliterated proper nouns) still beats the zero-overlap slice (n=7,555) by 6–13 Recall@1 points for every model.

Model	Form	R@1	R@10	R@100	NDCG@10
BGE-M3	Title	0.600	0.796	0.879	0.699
	Redirect	0.549	0.772	0.867	0.660
	Alias	0.503	0.749	0.861	0.626
	Desc.	0.120	0.306	0.612	0.203
mE5-large	Title	0.543	0.766	0.870	0.655
	Redirect	0.488	0.732	0.854	0.608
	Alias	0.454	0.712	0.840	0.579
	Desc.	0.085	0.262	0.602	0.162
Qwen3-Emb-4B	Title	0.745	0.868	0.909	0.810
	Redirect	0.686	0.841	0.894	0.769
	Alias	0.648	0.846	0.911	0.754
	Desc.	0.152	0.403	0.745	0.265

Table 4: Retrieval metrics by query formulation type. Query counts: Title=10,000, Redirect=4,277, Alias=1,035, Desc.=3,073.

Further breakdowns in Table 5, heatmaps in Appendix A, Figs. 3–4).

7 Do Hard Negatives Actually Confuse the Model More than Random Distractors?

This section highlights the methods and subsequent findings related to hard-negatives introduced in §4.

Method. For each of the 10,000 gold entities, we tag each top-10 retrieved document by `doc_type` and compare its observed share of top-10 slots against its corpus base share (gold 29.9%, hard negative 10.4%, random distractor 59.7%). We also check, per gold entity, whether mined hard negative outranks gold. We validate difficulty via model behavior rather than human judgment, consistent with prior model-agnostic hard-negative work.

Result. Random distractors appear in the top-10 results far less often than their share of the corpus (only 0.31–0.58$\times$ their base rate) for every model. Hard negatives are also under-represented, but much less severely (0.69–0.87$\times$ their base rate). Relative to corpus frequency, a hard negative is therefore roughly twice as likely to reach the top-10 as a random distractor. Gold documents are over-represented, as expected (lift 1.88–2.50$\times$; full numbers in Table 6, Appendix A, Fig. 4c–d).

Despite this increased difficulty, hard negatives almost never outrank the correct gold document at rank 1 (only 0.14–0.41% of cases; Appendix Fig. 5a). The same $\sim$2$\times$ relative-lift pattern holds across all three models, which use different pre-training data, objectives, and prompting conventions (§5). This consistency indicates that the hard negatives bring a genuine topical-confusability signal, in the same spirit as model-agnostic hard-negative validation in prior work [Xiong et al., 2021, Qu et al., 2021].

8 Conclusion

We presented a reproducible, fully-automatic pipeline for constructing English–Nepali cross-lingual retrieval corpora from public Wikipedia and Wikidata APIs, instantiated as WiNCC: 33,464 documents, 18,385 query formulations across 10,000 gold entities, and 3,482 class-matched hard negatives. Beyond construction, we introduced a metadata-sliced evaluation methodology that pinpoints exactly where cross-lingual difficulty concentrates (description-type queries, taxon-category entities), using three independently-trained multilingual embedding models to verify this. Results also show roughly double the retrieval lift of random distractors without ever outranking gold documents. Together these numbers support the corpus's use as a genuine difficulty-aware cross-lingual benchmark and the pipeline's use as a reusable template for other low-resource language pairs.

9 Limitations and Discussions

- **Structure-based labeling.** Language labels obtained faithfully from Wiki language site partitioning, and relevance is simplified to binary, not graded.
- **Rate-limiting affected hard-negative coverage.** Rate-limiting during Part 2 curation affected the extraction of queries and dropped the attainment to 19,982 distractors (§4 and Table 2) than the aimed 20,000 (as is in Part 1) .
- **Single language pair, non-comparative baselines.** Only English–Nepali language pair studied; §5 numbers are reference baselines only.
- **Comparison table.** Table 1 includes retrieval, bitext as well as QA onto shared axes for orientation only, not like-for-like equivalence comparison.

The class-matched hard negatives, proved in §7 to carry a genuine, and consistent confusability signal help in difficulty-controlled evaluation. The metadata-sliced methodology of §6 supports slice-based error analysis. Hard-negative pools can serve as contrastive training pairs [Karpukhin et al., 2020, Izacard et al., 2022], mined independently of any model's performance, unlike iterative model-dependent mining approaches [Xiong et al., 2021]. As Table 1 shows, WiNCC is, to our knowledge, the only automatic CLIR resource combining a dedicated Nepali subset, class-matched hard negatives, multiple query formulations per entity, as well as per-instance quality metadata simultaneously; the pipeline itself is reusable for other Wikipedia/Wikidata language pairs facing the same low-resource evaluation gap. The baselines and metrics are reference numbers attached to this release, not the paper's main claim, and are a cheaper but weaker form of evidence, complementary to annotation by humans supporting MIRACL and Mr. TyDi.

References

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. Xor qa: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 547–564.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4555–4567.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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 (ACL)*, pages 8440–8451.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research (TMLR)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5835–5847.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1351–1361.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the 7th Workshop on Challenges in the Management of Large Corpora (CMLC-7), 2019, Cardiff, Wales, UK*, pages 9–16.

Shuo Sun and Kevin Duh. 2020. Clirmatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4003–4012.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning (MRL)*, pages 127–137.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Extended Results and Visualizations

This appendix gives the full metadata-sliced breakdowns and supporting visualizations summarized in §6–§7, for all three baseline models, enlarged here for readability.

A.1 Query-Formulation-Type and Ambiguity Slices

Corresponds to the query-type and ambiguity discussion in §6. Figure 2(b)-(c) covers this.

A.2 Entity-Category and Lexical-Overlap Slices

Corresponds to the entity-category and lexical-overlap discussion in §6. Panels (a)–(b) of Figure 4 below cover this subsection; panels (c)–(d) are referenced again in §A.3.

A.3 Hard-Negative Validation

Corresponds to §7. Panels (c)–(d) of Figure 4 show the same top-10 composition and lift results tabulated exactly in Table 6 below.

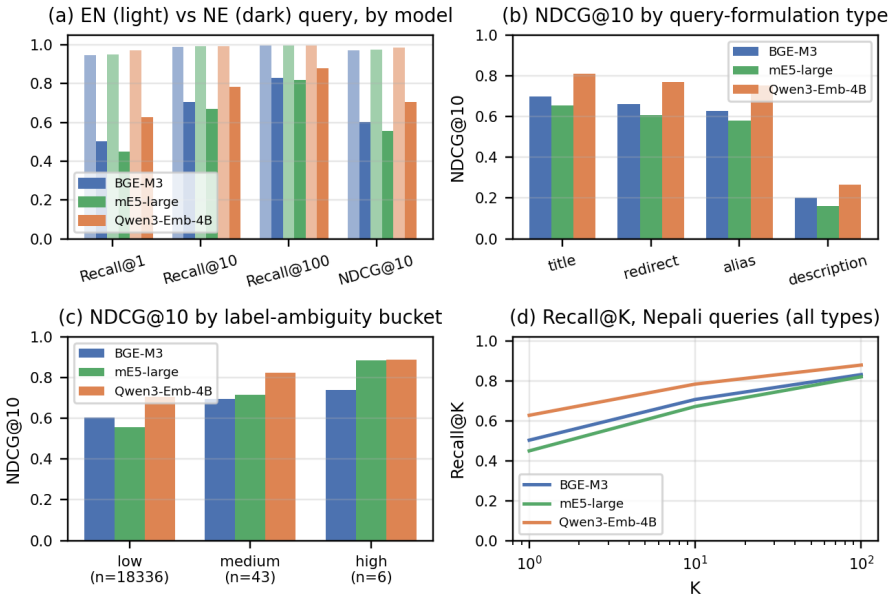


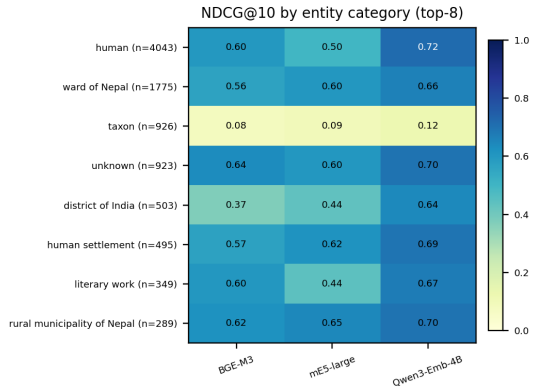
Figure 3: (a) EN (light) vs. NE (dark) pooled metrics by model; (b) NDCG@10 by query-formulation type; (c) NDCG@10 by ambiguity bucket; (d) Recall@K curve, Nepali queries (anchored at exact R@1/R@10/R@100).

Model	Metric	human	ward of NP	taxon	unknown	dist. of India	human settl.	lit. work
BGE-M3	R@1	0.502	0.467	0.054	0.536	0.239	0.465	0.513
	NDCG@10	0.599	0.564	0.081	0.635	0.371	0.568	0.604
mE5-large	R@1	0.388	0.518	0.064	0.504	0.322	0.511	0.335
	NDCG@10	0.498	0.603	0.094	0.604	0.435	0.615	0.443
Qwen3-Emb-4B	R@1	0.644	0.579	0.093	0.626	0.517	0.604	0.602
	NDCG@10	0.715	0.663	0.123	0.699	0.644	0.694	0.672

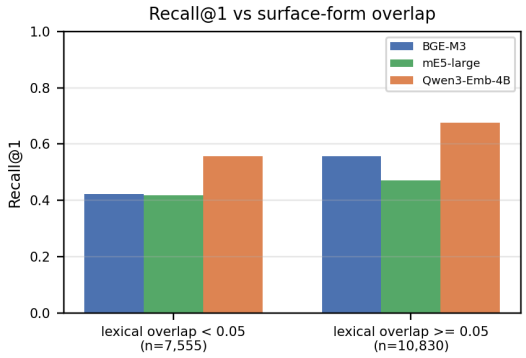
Table 5: Retrieval metrics sliced by entity category (top-7 of top-8 shown; `rural municipality of Nepal`, n=289, omitted for space – see Figure 4a for the full 8-category heatmap including it).

Model	Type	Obs. share	Base rate	Lift
BGE-M3	gold	0.563	0.299	1.88×
	hard neg.	0.091	0.104	0.87×
	random	0.346	0.597	0.58×
mE5-large	gold	0.746	0.299	2.50×
	hard neg.	0.072	0.104	0.69×
	random	0.182	0.597	0.31×
Qwen3-4B	gold	0.674	0.299	2.26×
	hard neg.	0.088	0.104	0.84×
	random	0.238	0.597	0.40×

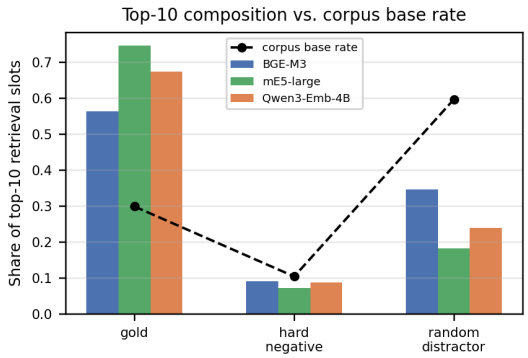
Table 6: Top-10 retrieval-slot composition vs. corpus base rate, by document type (title-query rows, n=10,000 per model). Lift = observed share / base rate.



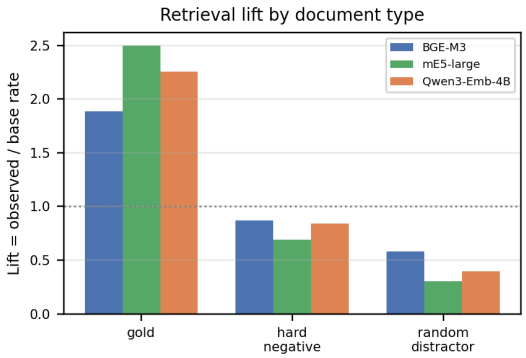
(a) NDCG@10 heatmap, top-8 entity categories × 3 models.



(b) Recall@1, zero- vs. some-lexical-overlap slices.

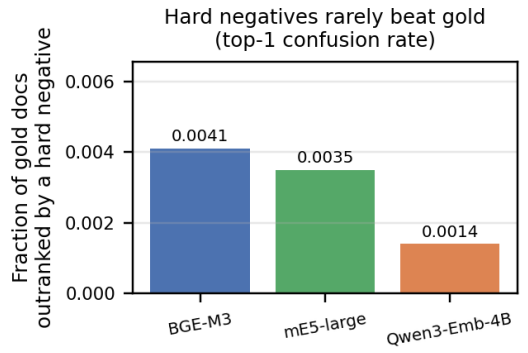


(c) Top-10 slot composition vs. corpus base rate.

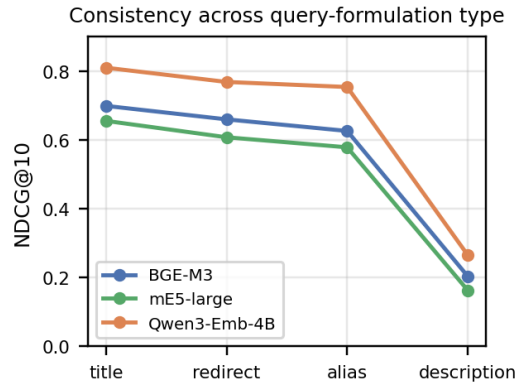


(d) Lift ratio (observed/base) by document type.

Figure 4: (a) NDCG@10 heatmap, top-8 entity categories × 3 models; (b) Recall@1, zero- vs. some-lexical-overlap slices; (c) top-10 slot composition vs. corpus base rate; (d) lift ratio (observed/base) by document type.



(a) Fraction of gold documents where a mined hard negative outranks gold at rank 1.



(b) NDCG@10 by query type, three models overlaid: an identical difficulty ordering (title > redirect > alias ≫ description) across architectures.

Figure 5: Hard-negative outranking rate (left) and cross-model query-type consistency (right).

Mapping Sámi languages in a Norwegian national web crawl: results from a representativeness study

Magnus Breder Birkenes
National Library of Norway
magnus.birkenes@nb.no

Abstract

The paper reports from a representativeness study covering Sámi languages in a Norwegian national web crawl. We extracted linguistic content from HTML in a considerable portion of the archive and applied language identification on paragraph-level. By aggregating the results on domain-name level, we were able to quantify the amount of Sámi language content in the web archive and identify contributing domains. Further, by comparing the URLs of the national crawl with a Sámi reference corpus, we identified domain names that were lacking in the national crawl and domains that lack depth. The results will be used to increase the coverage of Sámi language content in future web crawls and acquisitions.

1 Introduction

In Norway, Norwegian and Sámi (North Sámi, Lule Sámi and South Sámi) are recognized as "languages of equal worth" under the *Act on the Sámi Parliament and other Sámi legal matters*, Section 1-5).[1] Cultural heritage institutions in Norway collect and make available content in these languages for the general public and research. These collections also include web documents.

The National Library of Norway, a major cultural heritage institution in Norway, has crawled the Norwegian web since the late 1990s. The *Legal Deposit Act*[2] grants the institution the right to crawl the top-level domain (TLD) *.no* and web pages with relevance for Norwegian culture and society, including indigenous peoples and minorities.

But even if this web crawl (henceforth: "the national crawl") is substantial in size, we know little about the distribution of languages in the resulting corpus, especially the Sámi languages and other minority languages such as Kven and Romani/Romanes. One could hypothesize that the crawl is less complete for these communities due to various factors such as political priorities, unclear responsibilities, and a lack of knowledge.

The current paper reports on a representativeness study on Sámi languages within the national web crawl. It is carried out as part of the WebData research infrastructure,[3] funded by The Research Council of Norway[4] and hosted at The National Library of Norway. Among the partners is Giellatekno at UiT The Arctic University of Norway which contributes expertise in the field of Sámi language technology,

The study has two goals: Firstly, it aims to provide an overview over web domains with Sámi language content within the national crawl. Secondly, it aims to identify domains and websites that have not been captured in the national crawl and that should be harvested (if still online) or acquired in retrospective through other archivist institutions (for historical material).

We are not aware of previous language-oriented representativeness studies of national crawls, but e.g. Jauhiainen et al. (2020) look into building web corpora for minority languages and Dodge et al. (2021) investigate data provenance in large web corpora like CommonCrawl.

2 Data

2.1 Raw WARC files from the national web crawl

A web archive typically consists of so-called WARC files, a standard format for capturing traffic between a crawler and the servers it visits. A WARC file is simply a sequence of client requests and server responses, with some additional metadata such as the captured URL, the time of record-

[1] https://lovdata.no/dokument/NLE/lov/1987-06-12-56 (unofficial translation by Lovdata)

[2] https://lovdata.no/dokument/NL/lov/1989-06-09-32

[3] https://webdata.nb.no

[4] RCN project number: 350342

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 139–146
October 29, 2026 ©2026 Association for Computational Linguistics

ing and the content format (like HTML or PDF) in addition to the content itself.

In this paper, we process WARC files from the national crawl captured between 2005–2009, 2012–2014 and 2019–2025. The missing years were not readily available for processing at the time of the investigation because they had not yet been deduplicated or could not be simply extracted due to storage limitations. Still we consider the selected periods representative enough to tell something substantial about the archive.

Further, using metadata of the WARC files, we restrict the following study to HTML pages, the core object of the web. Other text formats such as PDF and DOC(X) were excluded due to restrictions of time. Especially the former typically needs more processing in order to get high quality text, such as layout analysis and OCR.

The raw data sizes of these HTML records per year are shown in Table 2 of Appendix A. We have processed a total of 1.8 billion HTML responses with 96 TB of payload. There are considerable differences from year to year, partly due to crawling intensity, partly due to data duplication.

2.2 Reference corpus

The most difficult part of the study is to answer what the national crawl does **not** capture. The web is enormous and dynamic, and there exists no such thing as a "complete" crawl. As an approximation, we choose to use the open corpora of the GiellaLT infrastructure[5] as a point of reference (Moshagen et al. 2023). We use the following Sámi corpora in this study: `corpus-sme-orig` (North Sámi), `corpus-sma-orig` (South Sámi) and `corpus-smj-orig` (Lule Sámi), as of March 2026.[6] For each corpus, we limit ourselves to HTML documents by only extracting the document URLs referencing actual HTML documents as per the bundled metadata XSL files. We stored the URLs referencing HTML documents in a flat text file.

3 Method

3.1 Data extraction

We process the raw WARC files using the Python package `warcio`,[7] filtering them to only include HTML content and responses with HTTP status code 200 (OK), which indicate that the web page existed at crawl time and was served successfully.

For the following study, we are only interested in the linguistic content of web pages. In addition to the markup itself, a typical web page consists of repetitive elements (headers, footers, navigation bars) and other elements that are not part of the main content of the page and are typically labelled as "boilerplate" (Bevendorff et al. 2023, 2594).

In this study, we use the boilerplate removal system `jusText` (Pomikálek 2011). `jusText` uses a combination of rules based on the semantics of HTML elements (e.g. `nav` and `footer` elements hint at repetitive content), link and stop word density to identify (contiguous) blocks of elements containing language. For the best possible extraction, we extended `jusText` with curated stop word lists for Norwegian and Sámi languages.[8] The output is simply a list of strings, one element per paragraph.

3.2 Language identification

Crucially, we run language detection on all extracted documents on paragraph-level, in order to account for multi-lingual texts. For language detection, we use a Python implementation[9] of the TextCat algorithm (Cavnar and Trenkle 1994) together with a set of publicly available models (including Norwegian and the Sámi languages) from GiellaLT.[10]

The TextCat algorithm computes character n-grams for each text, comparing the particular distribution with a model using out-of-place measurements. Since language detection systems in general struggle with very short texts (e.g. van der Goot 2025), we do not classify documents or paragraphs that are shorter than 10 whitespace separated tokens. All paragraphs above this threshold receive exactly one label.

3.3 Deduplication

In a web crawl, one URL is typically captured multiple times and even though the web is dynamic, often the main content remains unchanged between frequent captures. When we count instances of a

[5] https://giellalt.github.io/AboutGiellaLT.html

[6] We thereby made sure that no HTML documents had been added after 2025.

[7] https://github.com/webrecorder/warcio/

[8] The stop word lists are available here: https://github.com/NationalLibraryOfNorway/maalfrid_toolkit/tree/master/src/maalfrid_toolkit/models

[9] https://pypi.org/project/giellatect/, a packaged version of GiellaLT's `pytextcat` module (https://github.com/giellalt/CorpusTools).

[10] https://github.com/divvun/CorpusTools/tree/main/corpustools/lm

document, we do so only for unique documents within one and the same web domain (e.g. *nrk.no*). We do this by sha1-hashing each document as extracted from `jusText` and picking the first instance. This is a simple and at the same time naïve approach as one single change in the document (e.g. a changed character) leads to a new checksum and hence a new document.

3.4 Data aggregation

We choose to aggregate the language counts per **host**. According to RFC 1738[11], a host is the "[t]he fully qualified domain name of a network host", that may include subdomains (e.g. *tv.nrk.no*).

We think that the host is a more suitable identifier to aggregate over than the domain itself: e.g. the domain *kommune.no* is used by all municipalities in Norway, whereas its subdomains point to each individual municipality. If we chose the domain as aggregation level, smaller languages like Sámi would nearly disappear. On the other hand, municipality subdomains like *snasa.kommune.no* or *kafjord.kommune.no* are expected to have a high volume of Sámi content. For reasons of simplicity, we will refer to both *kommune.no* and *snasa.kommune.no* as domain names.

In order to filter out noise, we only consider content from domain names that have at least 500 tokens in any of the three considered Sámi languages.

3.5 URL set comparison

To assess the overlap between the national crawl and the reference corpus, we compare sets of normalized URLs in the two crawls. Importantly, we only look at the URL strings themselves, and do not investigate what the content looked like at crawl time or the language of the content.

We apply a set of rules to each captured unique URL in the two corpora: We 1) convert the URL to lowercase, 2) remove the protocol (http/https), 3) merge subdomains that begin with *www* with the main domain, since these two are often used interchangeably, 4) strip optional trailing slashes, 5) sort the URL parameters alphabetically, and finally 6) cut the string at the hash, effectively removing internal links and JavaScript directives.

Since the analyzed parts of the national crawl consist of 1.8 billion responses (cf. Table 2 of Appendix A), we first reduce the set of URLs from the national crawl to the domain names captured in the reference corpus before applying the above set of normalization rules. This step reduces computational costs and does not affect URL-level overlap because URLs from domains absent from the reference corpus cannot contribute to the overlap.

[11] https://datatracker.ietf.org/doc/html/rfc1738

lang	hosts	tokens	share
nob	327,123	29,951,618,479	68.16%
eng	131,526	10,611,419,013	24.15%
nno	45,726	2,019,595,019	4.60%
other	48,439	1,317,242,029	3.00%
sme	725	39,130,826	0.09%
sma	145	3,369,714	0.01%
smj	154	2,224,880	0.01%

Table 1: Overview of languages in the national crawl

4 Results

4.1 Overview of Sámi language content

Table 1 illustrates the extent of Sámi languages in the national crawl. As a point of comparison, we have included the two official Norwegian languages (Bokmål and Nynorsk) as well as English. Other languages were collapsed into a label *other*.[12]

In comparison to the other languages, the number of identified hosts containing Sámi content is rather low. Of a total of 43.9 billion tokens, the Sámi languages account for 44.7 million tokens or about 0.11 % of the archive. Still, the raw token counts for North, Lule and South Sámi align quite well with the reported corpus size for these languages, which was 38.18 million tokens in 2024 (cf. Wiechetek et al. 2024, 15926)

Table 3, 4 and 5 in Appendix B show the top 20 hosts in terms of token volume for each of the three languages. When looking at the individual hosts, two numbers are particularly interesting: the absolute number of tokens in a Sámi language and the relative amount of tokens in Sámi on that particular host.

Exploring tables across the identified 831 unique hosts with Sámi content can be challenging qualitatively. We therefore developed a web application for the project that visualizes these two dimensions and lets the researchers look at individual hosts and their contents qualitatively using an interactive scatter plot. A screenshot of the application is shown in Figure 1.

[12] We recognized content in 46 languages, but note that our language detection system only supports 51 languages.

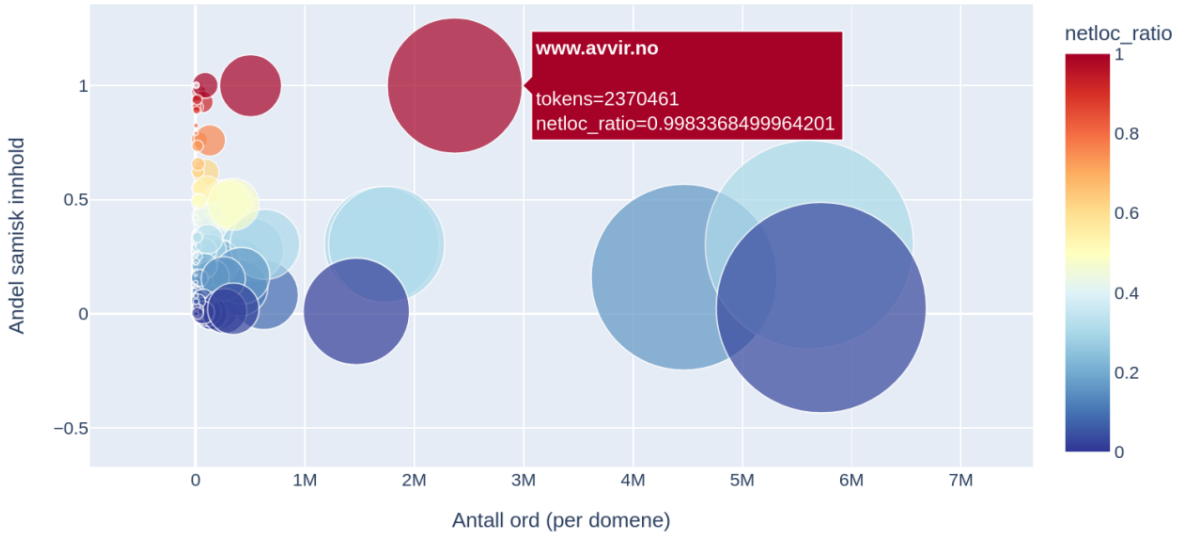


Figure 1: Screenshot of the web application used for qualitative inspection of relevant domain names and contents

The x-axis shows the number of tokens, the y-axis shows the relative amounts of Sámi tokens to other languages. The distributions of the two variables are heavily skewed: most hosts have a small amount of Sámi tokens, the median being at 2,820 tokens, but there is still a long tail of hosts with a very high amount of tokens (up to 6.7 million). Equally, the ratios concentrate around 0 %, with a median at 1 %.

There are two interesting and important outliers: The online newspaper Ávvir (*avvir.no*) publishes content nearly exclusively in North Sámi, at the same time having a large volume of tokens (2.3 million). The Norwegian Government (*regjeringen.no*), on the other hand, has a much lower proportion of Sámi languages (1.1 %), but due to the size of the web page, the volume (1.6 million tokens) is still substantial.

4.2 Comparison with a reference corpus

4.2.1 Overlap in domain names

In part two of the study, we use set comparison to analyze the two URL sets described in section 3.5. First we extract the domain names from the normalized URLs. Out of the 112 unique domain names in the reference corpus, the national crawl captured content for 80 of them (71.4 %). Of these 80 domain names, 50 are from the *.no* TLD. The remaining domain names are found among *.se* (14), *.fi* (6) and international TLDs like *.org* and *.com* (10 in total).

On the other hand, 32 domain names of the reference corpus were **not** captured by the national crawl. Only three of these domain names belong to the *.no* TLD, which is consistent with the national crawl setup: *.no* domains are typically accepted automatically, whereas other TLDs are subject to an individual assessment. 24 of the missing domain names are in neighboring Sweden (13) and Finland (11), whereas the remaining five are on other international TLDs. Most missed domain names belong to state and municipality-level institutions.

4.2.2 Crawl depth

We also investigate the depth of the national crawl by looking at the union of the two URL sets described in section 3.5. Selecting URLs from the 112 domain names of the reference corpus, the reference corpus contains 13,722 unique URLs and the national crawl 6,612,397 unique URLs after normalization. Importantly, both URL sets reference HTML documents only.

The result is that the national crawl only captures 19.7 % of URLs of the reference corpus. This tells us something important: Whereas the domain-level overlap was substantial, the URL-level intersection proved more limited. Even when reducing the URLs of reference corpus only to the overlapping domain names, the coverage increases only slightly, to 22.2 %.

Of the 112 domain names of the reference corpus, the national crawl has URL-level overlaps only with 35 of these. And only on 18 of these 35 domains have more than 50 % of the URLs of the reference corpus been captured. For the remaining domain names, the national crawl either captures nothing (32) or other URLs (45). Even within the

national administration, the coverage seems low, with only 11 % URL overlap on the website of the Norwegian government (*regjeringen.no*).

5 Discussion and outlook

Part 1 of the study identified a sizable amount of Sámi content (measured in tokens) in the national crawl. When compared with a reference corpus in part 2, we notice that even if a substantial portion of the relevant domain names has been visited by the crawler, the crawl seems to lack depth, as only about 1/5 of the URLs of the reference corpus has been captured by the crawler.

We want to stress that the numbers need to be interpreted cautiously. For certain domains like *nrk.no* and *bibel.no* we have identified a larger share of near-duplicates, which artificially increase the token counts. A future investigation should apply methods such as SimHash or MinHash to reduce the amount of such near-duplicates which are characteristic in web archives. The same cautiousness applies to the second part of the study: Comparison of URLs across harvests at different times is problematic since the the web is highly dynamic and URLs are subject to change.

The results of this study are important for the ongoing infrastructure project where the national crawl will be made available for research purposes: The resulting list of domain names can, after manual inspection, be used as seeds for the crawler. The identified lack of depth in existing domain names indicates that a more focused crawl should be conducted for the Sámi language domains in order to capture as much content as possible. For historical material, the data would need to be acquired through other web archives, e.g. from those maintained in the Nordic countries.

Ultimately, the goal of this endeavor is to increase the coverage of Sámi content in the web archive. This can both stimulate research on Sámi language and culture and the development of language technology for Sámi languages. We also hope that this study can serve as inspiration for similar initiatives at other cultural heritage institutions.

Ethical considerations

We have chosen not to publish the complete lists of hosts and URLs for reasons of privacy. The domain lists in the appendix are therefore limited to the top 20 hosts per language. Certain host names, such as those used in blogs, contain potentially identifiable personal information. The complete lists are available upon request with a valid research purpose.

Limitations

The paper considers only a part of the national crawl, with several years missing, and only HTML pages with text extracted by `jusText`. As noted in the discussion, the raw token numbers need to be interpreted carefully, since we did not apply near-duplicate detection. Further, and very importantly, the data have not yet been investigated manually by speakers with proficiency in Sámi languages, which will be essential in the process of increasing the coverage of Sámi languages in the web archive. The language detection system, while used extensively by GiellaLT and The National Library of Norway, has to our knowledge not been externally audited or evaluated. Finally, the study relies on URL set comparisons and not content comparison when comparing with a reference corpus due mainly to differences in harvesting.

Acknowledgments

We are grateful to Trond Trosterud for pointing us to the GiellaLT corpora and for comments on the token counts and domain distribution presented in this paper.

References

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*, SIGIR '23, page 2594–2603, New York, NY, USA. Association for Computing Machinery.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of third annual symposium on document analysis and information retrieval, Las Vegas, April 1994*, page 161–175.

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, Online and Punta Cana, Dominican Republic.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2020. Building web corpora for minority languages. In *Proceedings of the 12th Web as Corpus Workshop*, pages 23–32, Marseille, France. European Language Resources Association.

Sjur Nøsterbø Moshagen, Flammie Pirinen, Lene Antonsen, Børre Gaup, {Inga Lill Sigga} Mikkselsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain. 2023. *The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP*, pages 70–94. NEALT Monograph Series. University of Tartu, Estonia.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis. Masaryk University Brno.

Rob van der Goot. 2025. Identifying open challenges in language identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18207–18227, Vienna, Austria.

Linda Wiechetek, Flammie Pirinen, Maja Lisa Kappfjell, Trond Trosterud, Børre Gaup, and Sjur Nørstebø Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.

A Raw data analyzed per year

Table 2 shows the processed WARC data in the national crawl, i.e. HTML responses with status code 200. The table shows the number of such responses per year and the aggregated content length, i.e. the aggregated size of the response payloads.

year	responses	size (GB)
2005	199,444,309	5,914
2006	213,317,637	6,921
2007	194,967,792	6,794
2008	449,943,668	16,995
2009	156,183,539	4,273
2012	9,085,080	496
2013	50,400,889	2,748
2014	50,344,623	2,610
2019	6,122,214	465
2020	28,077,300	2,182
2021	17,854,548	1,540
2022	9,033,690	961
2023	124,720,421	11,910
2024	170,892,021	17,392
2025	130,324,074	14,933

Table 2: Processed HTML responses and size of HTML payloads per year in the national crawl

B Top 20 domains per Sámi language by token counts

Tables 3, 4 and 5 show the top 20 hosts in terms of token counts for North, Lule and South Sámi respectively. The data have been deduplicated at both the content and domain levels as described in Section 3.3. *tokens* refers to the aggregated token count of paragraphs classified as the given language, *paras* is the number of such paragraphs, *docs* is the number of documents with at least one paragraph in the respective language. *ratio* is the proportion of the respective Sámi-language tokens to all other tokens on that host.

host	tokens	docs	paras	ratio
www.nrk.no	5,722,732	36,028	222,976	2.70%
www.gavpi.org	5,614,019	34,292	189,227	30.24%
bibel.no	4,465,546	18,395	37,212	15.99%
www.avvir.no	2,370,461	11,751	70,926	99.83%
samisketallforteller.no	1,747,008	1,562	22,570	30.35%
samilogutmuitalit.no	1,701,250	1,513	21,899	30.48%
www.regjeringen.no	1,469,519	6,304	45,292	1.03%
www.nsr.no	634,133	3,226	16,181	30.45%
samiskbibliotektjeneste.tromsfylke.no	623,469	5,284	16,402	8.26%
skuvla.info	516,893	339	8,666	28.03%
www.minaigi.no	502,330	1,344	3,815	99.78%
www.samediggi.fi	418,638	1,128	10,759	16.63%
www.kafjord.kommune.no	403,705	3,303	11,526	11.39%
sametinget.no	388,392	1,729	11,034	20.90%
www.samediggi.no	345,324	1,542	6,003	47.93%
www.unn.no	344,787	8,208	15,003	2.19%
samas.no	312,118	1,706	9,040	46.84%
samiallaskuvla.no	289,667	1,740	8,233	48.53%
allaskuvla.no	287,486	1,727	8,172	48.21%
samiskhogskole.no	286,837	1,721	8,165	48.23%

Table 3: Top 20 hosts for North Sámi

host	tokens	docs	paras	ratio
www.nrk.no	587,232	3,528	24,741	0.28%
www.nord-salten.no	218,426	1,083	4,720	2.03%
www.gavpi.org	207,666	4,376	6,857	1.12%
www.udir.no	132,606	450	3,170	0.61%
www.infonuorra.no	106,087	507	2,865	4.09%
samiskbibliotektjeneste.tromsfylke.no	83,875	1,212	2,438	1.11%
issuu.com	80,003	978	1,737	0.00%
nordsalten.no	63,317	257	855	3.34%
www.regjeringen.no	54,628	271	1,579	0.04%
sametinget.no	40,592	274	1,044	2.18%
www.samedigge.no	39,366	173	1,051	6.71%
giellatekno.uit.no	31,979	11	1,517	4.83%
www.calliidlagadus.org	29,160	192	192	55.80%
www.statped.no	24,749	476	810	0.19%
nsr.no	24,722	58	736	3.07%
arran.no	23,215	125	679	18.94%
samiskeveivisere.no	19,944	76	587	3.88%
forum.infonuorra.no	19,094	202	248	2.42%
www.hamaroy.kommune.no	19,062	194	584	1.42%
sapmi.arbeiderpartiet.no	16,241	9	634	12.42%

Table 4: Top 20 hosts for Lule Sámi

host	tokens	docs	paras	ratio
www.nrk.no	407,652	2,817	17,338	0.19%
www.gavpi.org	339,209	6,848	12,357	1.83%
samisketallforteller.no	311,033	807	5,188	5.40%
samilogutmuitalit.no	299,630	786	4,970	5.37%
bibel.no	289,371	1,106	1,194	1.04%
www.infonuorra.no	281,593	1,377	9,034	10.85%
www.udir.no	208,112	554	5,286	0.96%
samiskbibliotektjeneste.tromsfylke.no	102,095	1,595	3,636	1.35%
www.regjeringen.no	92,655	657	2,594	0.07%
www.raadstua.no	84,800	1	684	50.46%
www.frisktigers.no	84,450	1,110	1,114	0.02%
sametinget.no	81,606	500	2,217	4.39%
issuu.com	78,694	593	1,183	0.00%
www.saemiedigkie.no	76,659	301	2,096	41.14%
www.statped.no	29,263	566	1,044	0.22%
www.statsforvalteren.no	26,124	206	798	0.11%
nsr.no	24,005	75	670	2.98%
ovttas.no	23,335	558	667	13.96%
www.kirken.no	21,615	126	632	0.11%
www.kyrkja.no	21,146	124	617	0.13%

Table 5: Top 20 hosts for South Sámi

Towards more reliable Web metadata: Do register classifiers learn register or genre features?

Antti Kanner, Tuomas Lundberg, Erik Henriksson, Veronika Laippala
TurkuNLP, University of Turku
antti.kanner@helsinki.fi
{tuomas.lundberg, erik.henriksson, veronika.laippala}@utu.fi

Abstract

Transformer-based web register classifiers are widely used to provide metadata for web crawls and to focus web-data sampling on specific register classes, such as news articles or advertisements. However, it remains unclear what these classifiers actually learn: do they detect pervasive linguistic features characteristic of registers, or do they rely on structural cues more indicative of genre? We examine this question through two hypotheses: the *text-macrostructure hypothesis* proposes that classifiers rely on structural features more typical of genres, whereas the *feature-density hypothesis* predicts that they identify typically register-associated features. Using English, Finnish, and French data from HPLT 3.0, we segment documents into thirds and compare classifier predictions for full documents and segments through five complementary analyses: main-label agreement, cosine similarity, confusion profiles, SACX keyword-weight correlations, and SACX keyword stability. As a more direct control, we compare predictions for a sample of untruncated documents with predictions for the same documents after their sentence order has been randomised. Top-label assignments and prediction vectors remain highly similar after shuffling, although the positional gradients are attenuated, indicating that sentence order contributes to classifier behaviour without determining its broader prediction profile. Overall, the results are more consistent with the feature-density hypothesis, although both hypotheses retain explanatory value.

1 Introduction

Web register and genre classifiers are already widely used to provide metadata for large web corpora such as HPLT and CLASSLA-web (Burchell et al., 2025; Pungeršek et al., 2026). This metadata supports the efficient use of heterogeneous web corpora by enabling informed sampling of their contents. Such sampling is beneficial in several applications, including the selection of training data for large language models (Myntti et al., 2025).

Automatic web-register identification has also advanced substantially in recent years. For example, the MultiCORE classifiers, developed using the Multilingual CORE corpora, achieve near-human performance (Henriksson et al., 2025). As these classifiers become more accurate and their outputs are used more widely, it becomes increasingly important to understand what they learn. From a practical perspective, this is necessary for evaluating the validity of the metadata they produce. From a theoretical perspective, automatic classification provides a new setting in which to examine the complex relationship between register and genre (Kuzman and Ljubešić, 2025; Laippala et al., 2022; Henriksson et al., 2025).

Register and genre are theoretically distinct notions, but they are not always explicitly defined in contrast to each other. Their boundaries also vary across theoretical frameworks. Consequently, their practical application may involve considerable analytical overlap, and register and genre categories may be strongly correlated in annotated data.

In this article, we approach this distinction by contrasting two hypotheses about the kinds of features on which a classifier bases its predictions. The first concerns pervasive features associated with registers, whereas the second concerns macrostructural properties associated with genres. We assume that a transformer-based classifier will exploit the most readily available structure in the data that corresponds to the categories on which it has been trained. Determining which kind of structure is more readily available to the classifier is therefore relevant not only to the present model but also to automatic text classification more broadly.

We therefore ask to what extent a multilingual web-register classifier relies on features that are pervasive across document positions, as opposed to features associated with the positional struc-

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 147–160
October 29, 2026 ©2026 Association for Computational Linguistics

ture of texts. We test two competing hypotheses about what the MultiCORE classifier learns. The *feature-density hypothesis* proposes that the classifier identifies varying densities of register-typical features distributed throughout texts. The *text-macrostructure hypothesis*, by contrast, proposes that the classifier relies on cues associated with text macrostructure and therefore focuses more strongly on genre-typical than register-typical features.

This experimental design cannot categorically determine whether the classifier learns registers or genres. Pervasive and macrostructural features do not map exclusively onto either concept, and the distinction between register and genre varies across theoretical frameworks. Instead, the analysis characterises the distributional organisation of the evidence used by the classifier. In doing so, it captures one commonly recognised distinction between registers and genres without treating that distinction as exhaustive.

Our analysis tests the two hypotheses by dividing each document into three segments—the beginning, middle, and end—and examining prediction consistency across these positions. We compare the predictions for each segment with those obtained for the corresponding full document. In addition, we analyse the stability of linguistic features across the three segments using SACX keywords (Rönnqvist et al., 2022). These keywords identify linguistic features that make important contributions to the classifier's predictions and therefore provide complementary evidence for evaluating the two hypotheses.

2 Background

In linguistics, register and genre are among the most common analytical frameworks for characterising and classifying texts and, more broadly, written communication (Lee, 2001; Biber and Conrad, 2019). The division of labour between the two concepts is not always clear-cut. Different theoretical frameworks define them in different terms, dispense with the distinction altogether, or favour one concept over the other.

One common distinction associates registers with functionally interpretable linguistic characteristics that are pervasive, meaning that they occur throughout a text. The sets of linguistic features associated with particular registers are often referred to as their repertoires (Biber and Conrad, 2019). Genres, by contrast, are commonly described in terms of their macrostructural organisation. This organisation may be characterised as a sequence of functionally distinct stages, sometimes—but not always—analysed as moves (Swales, 1990). The contrast is compatible with the text-linguistic approach presented by Biber and Conrad (2019), in which register refers to situationally and culturally defined varieties of language, such as news articles, advertisements, and student essays.

This theoretical distinction between register and genre raises a corresponding question about what automated register classifiers learn. Previous research has shown that register classifiers trained only on the beginnings of web documents can perform as well as classifiers trained on whole texts (Laippala et al., 2023). This result admits at least two interpretations. The classifiers may identify registers through features that occur pervasively across documents. Alternatively, they may learn cues associated with text structure, such as opening sequences characteristic of particular genres (Biber and Conrad, 2019; Swales, 1990; Dijk, 1980; Martin and Rose, 2005).

If register and genre categories correlate in the training data, a classifier may perform equally well regardless of which of these types of evidence it has learned to use. Predictive performance alone therefore does not reveal whether the classifier relies primarily on pervasive register-associated features or positionally organised macrostructural cues. From the perspective of classifier development, however, it is important to understand what kinds of features these models rely on and with which theoretical constructs their outputs align most closely.

3 Data and register classifier

For this study, three samples of 100,000 documents were drawn from the English, Finnish, and French components of HPLT 3.0 (Oepen et al., 2025). To focus the analysis on relatively coherent texts, sampling was restricted to documents with a Word Document Score (WDS) of 9 or 10. WDS is a document-quality measure based on features associated with suitability for language-model training, including linguistic consistency and the structural cohesion and integrity of the document (Oepen et al., 2025). Each document was retained as a full-document representation and additionally divided into approximately equal beginning, middle, and final thirds according to whitespace-delimited word count. The classifier was applied separately

to the full-document representation and each of the three segments.

We also constructed an additional control sample from the 100,000-document English sample to assess the effect of document truncation and test how classifier predictions changed when sentence order was randomised. Candidate documents were restricted to 1,000–1,500 whitespace-delimited words. This provided documents with sufficient internal extent to form three substantial segments while reducing the likelihood that they would exceed the classifier's 2,048-token input limit. Approximately 20,000 documents in the English sample met this word-count criterion, from which we randomly sampled 10,000. Token lengths were calculated using the `FacebookAI/xlm-roberta-large` tokenizer employed during prediction. Of the sampled documents, 658 exceeded the 2,048-token limit and were excluded. A further four documents were excluded because fewer than three sentences were detected, leaving a final paired control sample of 9,338 documents.

Sentence boundaries in the control sample were detected using the rule-based sentencizer in a blank English spaCy 3.8.16 pipeline (`spacy.blank("en")` with the `sentencizer` component). Each retained document was processed in two conditions. In the original-order condition, sentences were retained in their original sequence; their concatenation formed the full-document representation, and the sequence was divided into approximately equal beginning, middle, and final thirds according to sentence count. In the shuffled condition, the complete sentence sequence was randomly permuted once. The same permutation was used to construct both the shuffled full-document representation and its three successive thirds. The original and shuffled full documents therefore contained the same sentences and differed only in sentence order. Document identifiers and record order were preserved across conditions, enabling paired document-level comparisons.

All texts and segments were classified using the MultiCORE register classifier based on BGE-m3-retromae (Chen et al., 2024; Henriksson et al., 2025). BGE-m3-retromae is based on XLM-R (Conneau et al., 2020), has a maximum input length of 2,048 tokens, and has achieved high performance in register-identification tasks (Henriksson et al., 2025). The MultiCORE classifier performs multilabel classification using a simplified version of the CORE scheme (Egbert et al., 2015; Laippala et al., 2022), comprising nine main registers and 16 sub-registers. The main registers are broad functional categories such as Narrative, Informational Persuasion, and How-to/Instructions, whereas the sub-registers are more specific categories such as Narrative blog and Recipe. The classification guidelines are available at `http://turkunlp.github.io/register-annotation-docs`.

In the main sample, full-document representations exceeding the model's 2,048-token input limit were truncated during classification. This affected 15.5% of the English documents, 14.9% of the Finnish documents, and 12.8% of the French documents. Consequently, full-document predictions for these cases disproportionately represent document beginnings, which may increase agreement between the full-document and beginning-segment predictions. We account for this asymmetry when interpreting the main results. The English control sample addresses it directly because all retained full documents fall within the model's input limit.

A register label was assigned whenever its predicted probability exceeded the optimised threshold of 0.35. Because the task is multilabel, a document can receive more than one register label. Documents commonly receive both a main-register label, such as Narrative, and a more specific sub-register label, such as News article.

As an informal validation check, we compared the classifier's main-register predictions against manual annotations for 717 texts from the HPLT sample. The classifier achieved an accuracy of 63.6% and a macro-averaged F1 score of 52.0%. These figures are not directly comparable with the near-human performance reported by Henriksson et al. (2025), as the present evaluation was not constructed as a formal benchmark. It was used only to establish that the classifier captured a meaningful register signal in the present sample before its behaviour was examined comparatively. Accordingly, the analyses below characterise the behaviour of this classifier on the present HPLT samples rather than the behaviour of register classifiers generally.

4 Methods for hypothesis testing

Our primary question is whether the classifier relies predominantly on register-associated features that recur throughout texts or on features associated with their sequential macrostructural organisation. We refer to these alternatives as the

text-macrostructure hypothesis (1) and the feature-density hypothesis (2). The analyses do not provide a categorical test of whether the classifier learns genres or registers. Rather, they characterise the positional distribution of the evidence used by the classifier and assess which hypothesis provides the better account of the observed patterns.

The comparison assumes that beginnings, middle sections, and endings tend to perform different functions within the organisation of extended texts. A classifier relying strongly on sequential macrostructure should therefore exhibit positional differentiation in its predictions or attributed features. By contrast, if the classifier relies predominantly on features that recur throughout a text, its behaviour should remain comparatively stable across positions. This does not require complete positional invariance: the density and salience of pervasive features may vary over the course of a document even when the classifier relies on a broadly shared repertoire.

The beginning–middle–end division is used as a coarse, cross-register proxy for positional macrostructure. It does not identify genre-specific moves or functional stages. Manually annotating such units would require different structural schemes for research articles, narratives, recipes, advertisements, discussions, and the other categories represented in the data; units produced under these schemes would not necessarily be comparable across registers. Such an analysis would answer the narrower question of whether the classifier recognises predefined structural units within particular genres, rather than the present question of whether its predictions generally depend on sequential document organisation. The positional division therefore provides a theory-neutral basis for comparison across the heterogeneous sample. The design rests on the assumption that, if reliance on sequential macrostructure is sufficiently systematic to explain the classifier's register predictions across this sample, it should produce aggregate differences among broad document positions. Finer genre-specific effects may remain undetected when their locations vary or when they occur within individual thirds, and the conclusions are limited accordingly.

We evaluate the hypotheses using five complementary tests:

1. agreement between full-document and segment main-register labels;
2. cosine similarity between full-document and segment-level prediction vectors;
3. confusion patterns between full-document and segment-level main-register predictions;
4. Spearman correlation of SACX keyword weights across document segments; and
5. overlap of SACX keyword sets across document segments and full documents.

Label agreement and cosine similarity measure the consistency of segment-level predictions with the corresponding full-document prediction. Because main-sample full-document inputs exceeding 2,048 tokens are truncated, these comparisons use a reference that disproportionately represents document beginnings in 12.8–15.5% of cases. The English untruncated control sample is therefore used to repeat the main-label agreement and cosine-similarity analyses under original and shuffled sentence order. The confusion matrices provide a more detailed view of whether positional disagreement involves stable or position-specific relationships between register categories. Finally, the two SACX analyses examine the positional stability of the features attributed importance by the classifier.

Under the text-macrostructure hypothesis, we expect comparatively strong positional differentiation: lower agreement between full-document and segment predictions, position-specific confusion patterns, weaker correlations between attributed-feature rankings, and less overlap between keyword sets. Under the feature-density hypothesis, we expect prediction patterns to remain broadly consistent and attributed features to exhibit a substantial shared core across positions. These expectations, summarised in Table 1, are comparative interpretive criteria rather than categorical outcomes required by either hypothesis.

5 Results

5.1 Main-label agreement

This test measures how many of the highest predicted register labels assigned to the full documents are also assigned to the document segments. In order to focus the analysis on the broader, functional patterns and limit the number of comparisons, this test focuses on the nine main register labels.

Table 2 shows that the overlap of register labels assigned to full documents and to segment

Test	Text macrostructure	Feature density
Label agreement	Lower overlap and stronger positional differences	High overlap and modest positional differences
Cosine similarity	Lower similarity and stronger positional differences	High similarity and modest positional differences
Confusion patterns	Position-specific reorganisation	Broadly stable patterns
SACX correlation	Weaker correlations between feature rankings	Stronger correlations within the shared feature core
SACX stability	More position-specific keywords	Substantial keyword overlap across positions

Table 1: Expected patterns under the text-macrostructure and feature-density hypotheses. These expectations are comparative rather than categorical.

	Beginning	Middle	End
Finnish	78.49%	71.62%	69.54%
French	73.98%	70.00%	68.93%
English	77.98%	72.20%	69.61%

Table 2: Label overlap between full documents and segments. Values are means. 95% CI is attained with $\pm$ 0.3% for each item

labels ranges from 69.54–78.49%. The progression is monotonic: the overlap is the highest in the beginnings and decreases consistently towards the endings. The overlap with full documents and beginnings is exaggerated by the classifier truncation procedure as they might overlap fully in the case of longer documents. With all languages, the change is gradual and predictable, rather than abrupt. The results correspond better with hypothesis (2) than (1), as the differences are moderate and rather show gradual weakening of the register signal than markedly different profiles for each segment, even after taking into account the possible effect of truncation.

The main-label agreement analysis was repeated on the untruncated English control sample under original and shuffled sentence order (Table 3). In the original-order condition, agreement between full-document and segment predictions declined from 83.77% for beginnings to 75.90% for middle sections and 72.46% for endings. After sentence shuffling, the positional gradient was substantially

Condition	Beginning	Middle	End
Original order	83.77%	75.90%	72.46%
Shuffled order	85.51%	80.62%	79.99%

Table 3: Agreement between the highest-scoring register labels assigned to full documents and their corresponding segments in the untruncated English control sample (n = 9,338). All 95% bootstrap confidence intervals have margins below 0.94 percentage points.

attenuated: agreement was 85.51% for beginnings, 80.62% for middle sections, and 79.99% for endings. The difference between beginning and final segments therefore decreased from 11.31 to 5.52 percentage points. Direct comparison of the original and shuffled full documents showed that 85.80% retained the same highest-scoring main-register label (95% CI [85.05%, 86.51%]). Sentence order therefore affects classification in a minority of cases and contributes to the stronger positional gradient found in naturally ordered texts, but most full-document classifications remain unchanged when macrostructural sequencing is disrupted.

5.2 Cosine similarity

The distribution of cosine similarities between the prediction vectors of full documents versus each of the document segments tells how similar the prediction patterns are overall regardless of which class is assigned as the main label. Especially in documents where the classifier assigns near equal probability to two register candidates, relying only on the main-label agreement might mask considerable underlying similarity of the prediction vectors.

All nine cosine similarity distributions (three languages times three segments) showed high agreement with regards to the shape of their distributions. All nine follow clear power law curve with the histogram peak in high similarity end bin. Thus, reporting averages for brevity is grounded. These are given in Table 4. The table shows that mean cosine similarities are very high overall. Consistent, gradual but modest decrease along text progression is visible on all languages. Despite the greater textual overlap between document beginnings and the truncated full-document inputs, cosine similarity with full-document predictions was only slightly higher for beginnings than for middle and final segments.

The cosine-similarity analysis was also supplemented with the sentence-shuffling control and it produced the same overall pattern as the main la-

	Beginning	Middle	End
Finnish	0.932	0.905	0.897
French	0.919	0.901	0.897
English	0.941	0.921	0.912

Table 4: Mean cosine similarity by segment position.

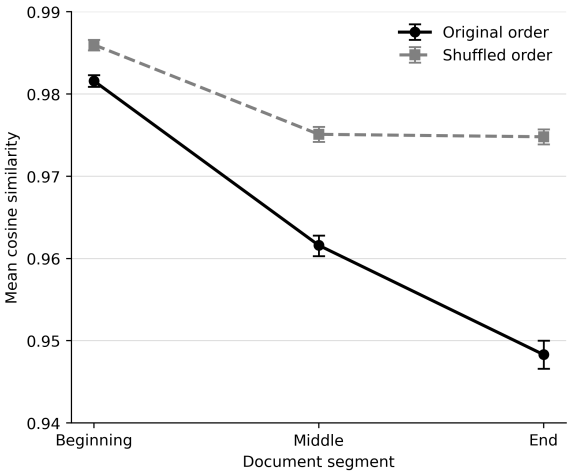


Figure 1: Mean cosine similarities between full-document prediction vectors and the beginning, middle and final thirds under the original and sentence-shuffled conditions. Error bars show 95% bootstrap confidence intervals with 2,000 resamplings.

bel agreement control analysis (Figure 1). The results also corroborate the patterns observed above. Reordering the sentences produced only a small change in the full-document prediction profiles: cosine similarity between the original and shuffled full-document vectors averaged .987 (95% CI [.986, .987]). Similarity between full-document and segment predictions also remained high in both conditions. In the original documents, however, mean similarity declined from .982 for beginnings to .962 for middle sections and .948 for endings. After sentences were shuffled and repartitioned, this positional gradient was substantially attenuated, with corresponding similarities of .986, .975 and .975. These results indicate that the classifier relies primarily on pervasive features distributed across the document, although naturally ordered texts appear to concentrate some diagnostically relevant evidence toward their beginnings.

5.3 Confusion profiles

Figure 2 reports the confusion profiles for each register in each segment in each language. The heatmap rows correspond to segment registers and

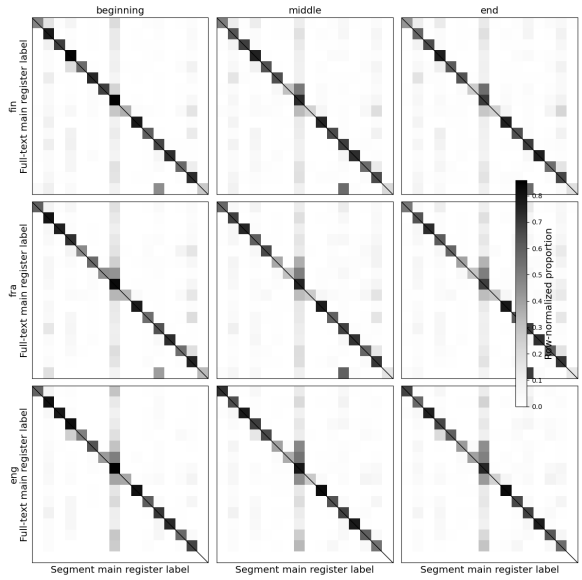


Figure 2: Confusion heatmaps for register predictions for full documents and segments.

columns to the full text registers.

The figure shows that in each heatmap, the diagonal line is strong, especially in the first registers. Some level of confusion is expected in the lower end for the sub-registers—these can be very detailed and easily confused, such as *encyclopedia article* and *description of a person or thing*.

All the heatmaps show strong vertical lines in the middle, in the same place. This is main register *opinion*, which captures registers like reviews, opinion blogs and advice pieces. The reason why this seems to be a strong fallback candidate for the classifier irrespective of language is difficult to assess. One explanation could be found from markers of evaluative language, which are crucial features of opinion registers, but can also be found in other registers (Koivunen et al., 2021).

Further, the figure clearly shows that there is no visible regime change in the later segments—the later segments are not systematically misinterpreted as texts belonging to other registers by virtue of their different structural properties. Quite the contrary, all segments show similar confusion profiles, with the gradual and modest loss of confidence visible as increase in confused mass, rather than its relocation. Further, there seems to be no major differences between the languages. All these findings suggest that feature density has the heavier impact on classifier behaviour.

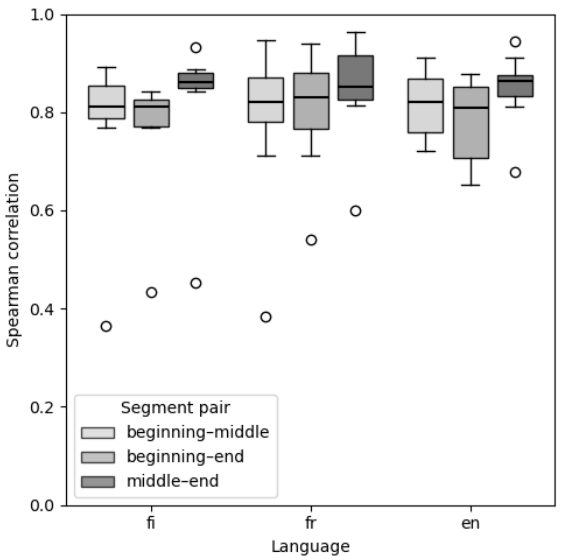


Figure 3: Spearman correlations of SACX feature weights between different text segments

5.4 SACX weight correlation across document segments

SACX (Rönnqvist et al., 2022) provides class-level keywords for text categories using the model explainability method Integrated Gradients (IG) (Sundararajan et al., 2017). For each document, and separately for each register predicted for it, IG accumulates the gradient of that register's score along a straight path in the embedding space, running from a baseline input in which every non-special token has been replaced by a filler token to the actual input, and assigns the resulting attribution to each token. The word-level attributions are then averaged over documents, yielding an importance weight for each word for each modeled class, where larger positive values indicate a stronger contribution to the prediction. In practice, this creates a list of keywords and weights for each modeled text category.

Our setup departs from the original SACX formulation in two respects: we use a single classifier rather than several independently trained ones, and the keyword lists are calculated with a 0.5 prediction threshold rather than the optimised 0.35 used for the label assignments reported elsewhere in this paper. Appendix A gives the details, along with the attribution settings.

For each language and main register, we selected words that occurred in at least 50 documents and had a positive mean attribution in all three segments. We ranked these shared-positive words by their mean SACX attribution separately in beginnings, middle sections, and endings, and calculated pairwise Spearman correlations between the three rankings.

The results of the analysis are presented in Figure 3. These show how similar or dissimilar the SACX weights are across the three document segments.

Overall, the correlations are very high. This means that the rankings of most important keywords across the different segments are essentially the same. The boxplots show the variation between the main registers. The low end outlier, especially prominent in Finnish and French, is Informational Description (IN).

5.5 SACX keyword stability across document segments

For this analysis, we used the keyword lists produced by SACX for each main register and language. The documents of each language and segment were processed in successive groups of 600, and a word was retained as a keyword for a register only if it was attributed to that register in at least 60% of those groups. This is the stability criterion for which the method is named: it keeps the words that are important across the whole set of documents and discards those that matter in only a few of them. We kept at most the 200 highest-scoring keywords per register. We call a keyword *pervasive* if it is selected in all three segments and *position-specific* if it is selected in exactly one. As an additional check, we compared the pervasive keywords against the keyword lists extracted from the full documents.

Table 5 reports the results. Averaged over the eight main registers with non-empty keyword lists, 42% (English), 40% (Finnish) and 46% (French) of the keywords are pervasive, and of these 71%, 65% and 60% respectively also occur among the full-document keywords. For the six registers whose segment lists are truncated at the 200-keyword limit these shares are lower bounds, since a keyword can fail the pervasiveness test simply by falling below the 200-word cut-off in one segment. The untruncated SP (Spoken) lists reach 73% (English) and 76% (French). The two analyses rank the registers differently — ID (Interactive Discussion) has the lowest pervasive share here, whereas the rank correlations place IN (Informational Description) lowest — which reflects the different sensitivities of set overlap and rank correlation rather than a

	English		Finnish		French	
Register	3/3	≥2/3	3/3	≥2/3	3/3	≥2/3
IN	40	59	29	56	40	62
NA	42	67	40	62	51	70
OP	47	69	44	70	44	67
HI	30	52	40	62	31	60
IP	41	66	50	67	46	69
MT	50	69	47	63	45	64
SP	73	89	55	82	76	86
ID	15	56	15	43	36	72
Mean	42	66	40	63	46	69

Table 5: Keyword pervasiveness % by language and main register. *3/3* and *≥2/3* denote the number of segments where the keywords were identified. For the first six registers, all three segment lists are truncated at the 200-keyword limit, so the 3/3 shares are lower bounds. The SP (Spoken) lists are untruncated in every language. LY (Lyrical) is omitted, as its keyword lists are empty.

substantive disagreement.

The position-specific keywords are distributed in the way the test setup anticipates. Excluding ID (Interactive Discussion), whose segment lists differ substantially in length, beginning-only keywords account for 39% (English), 42% (Finnish) and 40% (French) of all position-specific keywords, middle-only keywords for 30%, 27% and 27%, and end-only keywords for 32%, 32% and 34%. The middle sections thus contribute the fewest position-specific keywords in all three languages, which is consistent with the assumption that beginnings and endings carry more structural signal in a text compared to middle parts.

Relaxing the criterion from three segments to at least two changes the picture substantially. Keywords stable in at least two of the three segments account for 66% (English), 63% (Finnish) and 69% (French) of all keywords, so around two thirds of what the classifier relies on is shared by most of a document rather than being confined to one part of it.

The keywords steadily identified in exactly two segments are not, however, merely borderline cases of pervasiveness: disproportionately often, the segment they are missing from is the beginning. Restricting this count to the six registers whose segment lists are all truncated at the same 200-keyword limit, so that differences in list length cannot bias the comparison, the missing segment is the beginning for 45% of these keywords, the end for 32%, and the middle for only 23%. Middle and end sections thus share vocabulary more closely with each other than either does with beginnings, and beginnings carry additional distinctive material on top of a shared core.

All of these observations point at the same conclusion. The text macrostructure hypothesis would predict that each part of a document signals its register by its own distinct means. What we see instead in our segment comparison is that there is a substantial shared core running through the whole document, with beginnings carrying additional material, being structurally the most marked part.

Qualitatively, the pervasive keywords are register vocabulary. For NA (Narrative), the top keywords contain clear register markers such as *blog*, *post*, *said*, *yesterday*, *comment*, *readers*, *news*, and *announced*, together with weekday names; for HI (How-to / Instructions), words like *recipe*, *tips*, *instructions*, *guide*, *steps*, *method*, and *follow*; IP (Informational Persuasion) includes *clients*, *product*, *prices*, *customers*, *affordable*, *offer* and *guarantee*; SP (Spoken) has *tell*, *yes*, *talk*, *say*, *feel*, *think*, and *yeah*. The position-specific keywords are by contrast recognisably macrostructural: the end-only keywords of IN (Informational Description) include *disclaimer*, *conclusion* and *methodology*, those of IP (Informational Persuasion) *consultation*, *appointment*, *booking* and *pricing*, and those of HI (How-to / Instructions) *finally*, *advice* and *tip*. The keywords named here are illustrative; Appendix B gives the ten highest-scoring keywords of each class for every main register and all three languages, without applying manual selection.

The above results provide the linguistic counterpart to the quantitative results presented in Sections 5.1–5.5, though the picture here is more mixed than in those sections. Under the strict criterion requiring a keyword to be selected in all three segments, a minority of retained keywords (40–46%) qualify as pervasive, and more than half are therefore position-specific to some degree — a pattern that on its own would be at least as consistent with the text-macrostructure hypothesis as with feature density. Two further observations tip the balance towards feature density. First, relaxing the criterion to keywords stable in at least two of three segments raises the shared share to 63–69%, indicating that most position-specific keywords under the strict criterion are missing from only one segment rather than confined to a single position — closer to noise at the margins of a shared core than to distinct per-position vocabularies. Second, the qualitative content of the keywords is asym-

metric in a way the text-macrostructure hypothesis does not predict: pervasive keywords are overwhelmingly register vocabulary (blog, recipe, tips, clients, product), while genuinely structural markers (disclaimer, conclusion, booking, pricing) are concentrated in the smaller position-specific set, and disproportionately at document endings. Taken together, these results support the feature density hypothesis as the better fit to the SACX evidence, though the strict-criterion overlap figures alone do not settle the question.

6 Conclusions

The results are more consistent overall with the feature-density hypothesis (2) than with the text-macrostructure hypothesis (1), although both account for part of the observed behaviour. Across English, Finnish, and French, the classifier's predictions remained broadly similar across document positions, while agreement with full-document predictions declined gradually from beginnings to middle and final segments. This pattern suggests attenuation of a shared register signal rather than a categorical change in the kinds of register assigned to different parts of a document.

The prediction-based analyses supported this interpretation. Main-label agreement and cosine similarity declined progressively over the course of documents, but cosine similarities remained high throughout. Confusion profiles were also broadly stable across positions: disagreement increased in later segments, but the alternative labels followed similar distributions rather than showing systematic position-specific reorganisation. These findings indicate that document position affects prediction consistency without producing a general shift to qualitatively different register profiles.

The SACX analyses produced a more mixed but compatible pattern at the feature level. Keyword rankings were strongly correlated across positions, indicating a stable core of features attributed importance by the classifier. Under the strictest criterion, only 40–46% of retained keyword types occurred in all three segments, a result that alone is at least as consistent with positional macrostructure as with pervasive feature density. However, 63–69% occurred in at least two segments, and the qualitative distribution was asymmetric: pervasive keywords predominantly represented register-associated vocabulary, whereas recognisably structural markers were concentrated in the smaller position-specific set, particularly at document endings. The SACX evidence therefore indicates a substantial shared feature core alongside genuine positional variation.

The sentence-order control provided a more direct test of the role of sequential organisation. In the untruncated English control sample, the same highest-scoring main-register label was assigned to the original and shuffled full documents in 85.80% of cases, while their prediction vectors had a mean cosine similarity of .987. Shuffling also reduced the beginning-to-end gradient in main-label agreement from 11.31 to 5.52 percentage points. Cosine similarity displayed the same attenuation: similarities between full documents and their beginning, middle, and final segments were .981, .961, and .948 in original order, compared with .986, .975, and .975 after shuffling. Sentence order therefore contributes to positional differences and changes the top prediction in a minority of documents, but disrupting document-level sequencing generally leaves the classifier's broader prediction profile intact.

Taken together, the results suggest that the MultiCORE classifier relies primarily on register-associated evidence distributed across documents while also exploiting positional and structural cues where these are available. The findings do not establish that register classifiers generally learn register rather than genre, nor do they exclude finer genre-specific structural effects that are not captured by the three-way positional division. They instead characterise the behaviour of this classifier on the present HPLT samples: its decisions are predominantly supported by a recurring feature repertoire, with sequential macrostructure providing a secondary contribution.

Limitations

The study operationalises sequential macrostructure through a three-way positional division rather than through annotation of genre-specific moves or functional stages. This makes comparison possible across heterogeneous register categories without requiring a separate structural framework for each category, but it cannot detect structural variation occurring within individual thirds or effects whose positions vary substantially across documents. The conclusions therefore concern broad positional organisation rather than all possible forms of genre structure. A study using manually annotated structural units would address the complementary but

narrower question of whether the classifier recognises predefined stages within particular genres.

In the main multilingual sample, full-document inputs exceeding 2,048 tokens were truncated during classification. This affected 15.5% of the English documents, 14.9% of the Finnish documents, and 12.8% of the French documents, causing their full-document predictions to represent document beginnings disproportionately. Comparisons with full-document predictions may therefore favour beginning segments. The untruncated English control sample addresses this confound directly and reproduces the overall pattern, but it does not remove the asymmetry from the main multilingual analysis.

The sentence-order control is also limited to English documents containing 1,000–1,500 whitespace-delimited words and meeting the study's high document-quality criterion. Its results may not generalise fully to shorter, noisier, or non-English web documents. Furthermore, the classifier achieved an accuracy of 63.6% and a macro-averaged F1 score of 52.0% against the available manual annotations. The study consequently characterises the distribution and stability of the evidence used by this particular classifier; it does not establish that its predictions are correct or that other register classifiers rely on the same features.

Acknowledgements

This work was supported by the Research Council of Finland through several projects: FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union – NextGenerationEU instrument), "Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora" (project 362459), and the "Centre of Excellence Human Diversity through Contacts" (project 374223). We also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 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*, volume 1: Long papers, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

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, Online. Association for Computational Linguistics.

Teun Adrianus van. Dijk. 1980. *Macrostructures : an interdisciplinary study of global structures in discourse, interaction, and cognition / Teun A. van Dijk.* Lawrence Erlbaum Associates, Hillsdale (N.J.).

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2025. Automatic register identification for the open web using multilingual deep learning. *Preprint*, arXiv:2406.19892.

Anu Koivunen, Antti Kanner, Maciej Janicki, Auli Harju, Julius Hokkanen, and Eetu Mäkelä. 2021. Emotive, evaluative, epistemic: A linguistic analysis of affectivity in news journalism. *Journalism*, 22(5):1190–1206.

Taja Kuzman and Nikola Ljubešić. 2025. Automatic genre identification: A survey. *Language Resources and Evaluation*, 59(1):537–570.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on*

Noisy User-generated Text (W-NUT 2022), pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

David Y. W. Lee. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72.

J. R. Martin and David. Rose. 2005. *Genre relations : mapping culture*. Equinox textbooks and surveys in linguistics. Equinox Pub, London ;.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. Register always matters: Analysis of llm pretraining data through the lens of language variation. In *Proceedings of the Second Conference on Language Modeling (COLM 2025)*.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *Preprint*, arXiv:2511.01066.

Taja Kuzman Pungeršek, Peter Rupnik, Vít Suchomel, and Nikola Ljubešić. 2026. The Growing Gains and Pains of Iterative Web Corpora Crawling: Insights from South Slavic CLASSLA-web 2.0 Corpora.

Samuel Rönnqvist, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter, and Veronika Laippala. 2022. Explaining classes through stable word attributions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1063–1074, Dublin, Ireland. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

John Swales. 1990. *Genre analysis : English in academic and research setting / John M. Swales.* Cambridge applied linguistics series. Cambridge University Press, Cambridge.

A SACX configuration

Attributions were computed at the embedding layer using ten integration steps, summed over the embedding dimensions, and normalised per document. Sub-word attributions were merged to the word level by retaining the largest absolute value within each word.

Two properties of the setup should be taken into account when comparing our results to other applications of SACX. First, in its original formulation, SACX assesses the stability of the attributions across several independently trained classifiers. Here, we use a single classifier, and stability is instead assessed with respect to variation across the document groups described in Section 5.5. The criterion therefore controls for document-level rather than model-level variation. Second, attributions were computed for every register whose prediction probability exceeded 0.5, which is stricter than the optimised threshold of 0.35 used for the label assignments reported elsewhere in this paper. In other words, the keyword lists are created using more confident predictions.

B Keyword lists

Tables 6–8 give the ten highest-scoring keywords in each of the three classes discussed in Section 5.5 — pervasive, beginning-only, and end-only — for every main register and each of the three languages. Keywords are ordered by mean SACX importance and no manual selection was applied, so these lists show the classes as the method returns them.

Register	Pervasive (all three segments)	Beginning-only	End-only
IN	abstract, appointed, references, employed, article, characters, authors, writers, allowing, diagnosis	episode, here's, contents, nowadays, discusses, phenomenon, definitions, bound, entrepreneurs, excellence	therapies, disclaimer, yes, litigation, gov, conclusion, html, legacy, associates, mentioned
NA	blog, post, said, yesterday, comment, writer, added, comments, photos, readers	blogger, blogging, blogs, writers, jan, week's, posting, appointed, bbc, baseball	concluded, bridge, scene, crew, videos, media, johnson, highlight, mike, match
OP	review, reviews, blog, post, writers, writer, album, episode, book, reader	yesterday, october, april, november, folks, recent, morning, dear, guide, writing	rating, amen, corruption, wise, tips, authors, drama, reminds, magazine, artist
HI	recipe, tips, instructions, guide, note, steps, step, method, try, ways	we, found, getting, home, taking, making, better, i, give, few	advice, tip, ingredients, finally, techniques, mix, helpful, useful, fun, carefully
IP	book, today, contact, clients, product, prices, price, customers, affordable, client	author, sale, stock, incredible, we've, course, recommend, days, reasons, master	consultation, appointment, specialist, booking, guaranteed, pricing, we'll, packages, package, proven
MT	winning, customer, mobile, video, device, app, post, website, software, quality	via, requirements, improve, penis, sort, betting, training, single, latest, choose	user, solution, enterprise, policy, european, china, version, collection, difference, digital
SP	tell, yes, you, talk, what, say, your, feel, us, experience	career, today, her, she	thank, next, future, best, day
ID	question, think, i'm, sure, help, like, am, really, my, i	been, all, even, time, for, be, how, out, from, make	luck, yeah, pm, advice, wow, questions, hey, game, personally, comments

Table 6: English: the ten highest-scoring keywords in each class, per main register, ordered by mean SACX importance. No manual selection was applied. LY (Lyrical) is omitted, as its keyword lists are empty.

Register	Pervasive (all three segments)	Beginning-only	End-only
IN	syntynyt, yhteystiedot, tietojen, helsingissä, henkilötietoja, nimellä, yksikkö, vaihtelee, syntyi, suhde	properly, tausta, ilmiö, ominaisuuksien, function, opt-out, as, lasketaan, necessary, peräisin	tehtävästä, etc, functionalities, henkilötiedot, evästeitä, advertisement, julkaisuja, osoite, lisätietoa, yrityksille
NA	toteaa, ottelussa, sanoo, kertoo, ottelua, kommentoi, ajattelin, ottelun, johtaja, blogissa	otteluun, haastattelussa, maanantai, ihmettelin, kuvittelin, kansanedustaja, törmäsin, lauantai, sunnuntai, käsiteltiin	kommentti, uutishuoneessa, blogspot, kirjoittaja, suunnitelmia, kommentit, postauksia, yötä, instagramissa, kerrottiin
OP	suosittelen, elokuva, raamatun, elokuvan, jeesus, teksti, esimerkki, jumala, uskon, kirjassa	viikolla, viikon, sanotaan, luin, näkökulmasta, perustuu, kävi, kertaa, olemme, toisin	kommentti, suositella, kommentit, kirjoittaja, blogin, kommentoi, ps, hauska, lähetä, lukija
HI	ohjeet, sekoita, vinkkejä, dl, g, kannattaa, ohje, minuuttia, muista, mikäli	artikkelissa, tällä, selvittää, läpi, oma, tietää, tapa, pois, tehdä, olet	leikkaa, tutustu, löytyvät, kokeile, lisätietoja, kerro, paista, huomioida, ruokaa, seuraavan
IP	kysy, tutustu, asiakkaan, löydät, tarjoamme, hinta, suomalainen, saadaan, tervetuloa, tuote	sijaitseva, vaihtoehdon, hotelli, kerromme, käytämme, oletko, yhdistyvät, materiaalit, upeat, vastaa	kanssamme, helsingissä, ammattilainen, asiakkaillemme, mieltä, lisätietoa, asiakkaidemme, puh, suomeen, yrityksesi
MT	ilmoitus, yleens, kumppani, sovellus, valikoima, kokemus, pelejä, suunnitelma, pelit, tuhansia	osoite, merkkejä, sarjakuva, suurta, elää, kaupunki, talo, xvideos, kuva, club	valikoiman, palveluiden, kyky, mahdollisuudet, ssa, tekniikka, ohjelman, teidän, tietokoneen, arvostelu
SP	miten, olen, mitä, nyt, niin, on, paljon, kanssa, kun, oli	—	en
ID	kiitos, hei, tuossa, mieltä, minulla, tosiaan, juttu, d, takia, varmaan	tähän, minun, vuotta, täällä, lisäksi, tuli, silloin, viime, heti, tullut	viestiä, haluaisin, moikka, kysyä, moi, muistaakseni, aiheesta, toivon, lähetä, lienee

Table 7: Finnish: the ten highest-scoring keywords in each class, per main register, ordered by mean SACX importance. No manual selection was applied. LY (Lyrical) is omitted, as its keyword lists are empty.

Register	Pervasive (all three segments)	Beginning-only	End-only
IN	étude, organisme, espèces, personnage, article, missions, l'étude, appelée, document, caractéristiques	née, l'organisme, agit, appareil, chimique, appelés, appelées, trouverez, présentation, table	prenant, oui, notes, essentiels, références, source, conclusion, résumé, consultez, améliorer
NA	blog, souligne, jeudi, communiqué, interview, explique, expliqué, annoncé, déclare, mercredi	organise, news, post, accueilli, exposition, spectateurs, ouvre, profite, l'annonce, artiste	espère, thierry, conclut, remercier, emplois, ailleurs, document, auteur, professeur, champion
OP	album, avis, blog, roman, commentaire, commentaires, recommande, série, message, critiques	croire, marque, propose, vais, joueurs, sérieux, parle, utile, textes, joie	conseille, l'auteur, étude, avantage, écouter, témoignage, avenir, apprécié, facebook, approche
HI	recette, article, conseils, procéder, ingrédients, voici, utilisez, méthode, étape, recommandé	modifier, allons, ici, éléments, permet, connaître, correctement, lors, fonction, quelques	recommandons, conseil, n'oubliez, formulaire, aidera, alimentation, obligatoire, commentaires, mesures, mettez
IP	ouvrage, proposons, exposition, découvrez, avantages, contactez-nous, roman, retrouvez, accompagnement, contacter	auteur, lecteurs, innovant, accueil, article, innovants, bienvenue, invitons, paradis, l'esthétique	newsletter, acteur, logo, commentaire, l'agence, témoignages, l'opportunité, spectacles, dédiée, promo
MT	utilisateur, fonctionnalité, avantage, tenant, appareil, website, appareils, utilisation, objectif, emploi	étoiles, confortable, sentir, décoration, note, telecharger, maison, population, basé, essentiel	alimentation, more, achats, assurance, efficacité, microgaming, écran, exercice, privée, augmentation
SP	oui, comment, question, votre, quel, projet, ans, avez, premier, france	interview, questions, directeur, président, politique, pendant, vie	projets, quels, niveau, mettre, personnes, sera
ID	question, bonjour, merci, mon, j'ai, j'ai, me, suis, par, depuis	—	forum, répondre, sujet, avis, semble, effet, plutôt, sinon, ta, france

Table 8: French: the ten highest-scoring keywords in each class, per main register, ordered by mean SACX importance. No manual selection was applied. LY (Lyrical) is omitted, as its keyword lists are empty.

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 Crawl[1] 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

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 161–169
October 29, 2026 ©2026 Association for Computational Linguistics

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.

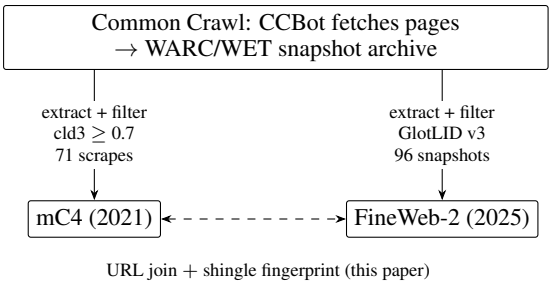


Figure 1: The two corpora are siblings rather than a derivation chain. Each selects and extracts documents independently from the shared Common Crawl archive, under its own LID filter, over snapshot ranges that overlap. Neither corpus performs its own crawl, and neither reads the other, so a document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

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 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 cld3[2] 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.176[3], 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.

	Shared with mC4		Flag rate, % (dual-classifier)			
	% of FineWeb-2 docs		mC4	FineWeb-2		
Language	co-sel.	page-sh.	all	co-sel.	unmatch.	$n_{\text{co-sel.}}$
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: refiltering 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. MC^2: 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`	*Suku sunda adalah salah satu suku yang ada di Indonesia. Ciri khas dari suku sunda adalah orangnya yang ramah da…* 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`	*Команда кафедры «Телематика» Кыргызского государственного технического университета им.И.Раззакова и ОФ «Технолэ…* *Komanda kafedry «Telematika» Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta im. I. Razzakova i OF «Tekhnole…* 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`	*Таразский государственный педагогический институт, г. Тараз В физкультурно-спортивной деятельности, как обоснова…* *Tarazskiy gosudarstvennyy pedagogicheskiy institut, g. Taraz V fizkul'turno-sportivnoy deyatel'nosti, kak obosnova…* Russian: 'Taraz State Pedagogical Institute, Taraz. In physical-education and sports activity, as substantiated…'
Yoruba	es / `spa_Latn` / `spa_Latn`	*Nueva Versión Internacional (Castilian) Todavía no hay audio para esta traducción. Un destino común 1A todo esto…* 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`	*Applications are NOW open for JUNIOR EUROVISION 2020 Do you have a great singing voice? Are you between the ages…*
Azerbaijani	en / `eng_Latn` / `eng_Latn`	*Arsenal leaves out Mkhitaryan from Europa League final amid safety concerns Armenian national Henrikh Mkhitaryan…*

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.

	mC4, all documents (%)					FineWeb-2, unmatched (%)				
Language	dual	sym	fastText	GlotLID	script	dual	sym	fastText	GlotLID	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∧GlotLID), symmetric (fastText∧OpenLID), fastText alone, GlotLID 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 GlotLID alone shows what requiring fastText's agreement costs: 2.4 points for Cebuano on mC4, under 0.8 points everywhere else.

	Co-selected (%)			Page-shared (%)		
Language	dual	sym	n	dual	sym	n
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.

How Many Duplicates Are Left? Capture–Recapture and Sampling Inversion for Residual Duplication in FineWeb

Gary Zhang
William A. Shine Great Neck South High School
zhanggary607@gmail.com

Abstract

Pretraining corpus releases report how much duplication their pipelines removed, but we are not aware of one that estimates how much they missed. Measuring the residue is difficult because two recall problems apply at once: detectors miss duplicates they cannot see, and any tractable subsample misses duplicates whose partners were not sampled. In this paper we address both. We 1) run four detectors with distinct failure modes over a corpus and treat them as capture occasions, 2) recover what all four missed through capture–recapture, and 3) invert the sampling rate that produced the subsample, so that the result describes a full 15-trillion-token corpus rather than a convenient slice of it. The design is scored first against a planted population of known size, where it recovers 89.2% of the truth from a 44.5% detected rate while remaining a lower bound. Applied to the 14,868,862-document `sample-10BT` release of FineWeb, 4.50% of probes have a surviving near-duplicate, and capture–recapture raises that to 4.81% (95% CI 4.37–5.26). Inverting a sampling rate of one document in 1,532 then carries the figure to the corpus and gives a lower bound of 68 on the mean near-duplicate degree of a FineWeb document. Near-duplicate here means lexical overlap at Jaccard 0.35 or above. The bound is on a mean and not on a prevalence, since our design leaves the share of documents with any surviving duplicate anywhere between 7% and 73%. The survivors also carry a positional signature. The two copies of an exact-duplicate pair sit a mean 34.4% of the corpus apart against 33.3% expected under uniform sampling, which supports reading them as residue of FineWeb's deliberate per-snapshot deduplication. A deduplicated corpus, therefore, is not a duplicate-free corpus, and the size of the gap can now be measured in one streaming pass.

1 Introduction

Deduplication is a standard stage of pretraining-data pipelines, because duplicated text distorts training. Duplicates drive memorization (Carlini et al., 2023), they inflate effective epochs on repeated content (Lee et al., 2022), and they leak evaluation data into training (Jacovi et al., 2023). Typically, corpus reports quantify the stage by what it removed. FineWeb, RefinedWeb and Dolma each describe their pipelines and the fraction of text discarded (Penedo et al., 2024, 2023; Soldaini et al., 2024), and readers are left to assume that the remainder is clean. No release we are aware of estimates the duplication that survived. The claim that a corpus is deduplicated is therefore unfalsifiable in practice.

Two distinct recall problems stand between a corpus and a correct residual number, and prior work addresses neither. The first is detector recall. Any similarity detector has a threshold, and duplicates below it are invisible, so reporting what one detector found says nothing about what all detectors would find. The second is observation recall, and it has, to our knowledge, gone unnoticed. Corpora at web scale are studied through subsamples, but duplication is a pairwise property, so a subsample destroys it quadratically: a sample retaining a fraction p of documents retains only about p^2 of the duplicate pairs. A duplication rate measured on a subsample and reported as a property of the corpus is thus too low by a factor of roughly $1/p$, which at the sampling rates web-scale work actually uses means orders of magnitude.

We solve both problems in one design. For detector recall we borrow capture–recapture, the estimator that ecology and census science use when detection is known to be incomplete (Chao, 1987; Fienberg, 1972; Hook and Regal, 1995). Each duplicate detector becomes a capture occasion, each document a unit, and the cell of documents that

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 170–181
October 29, 2026 ©2026 Association for Computational Linguistics

every detector misses is what the estimator recovers. For observation recall we invert the sampling rate analytically. The rate at which sampled documents find a partner inside the sample is a first moment of the population degree distribution, so dividing by p recovers the mean number of duplicates a document has in the whole corpus, without any distributional assumption.

The design is applied to FineWeb (Penedo et al., 2024), whose `sample-10BT` release is small enough to index completely and popular enough that its properties reach many training runs. We 1) fingerprint all 14,868,862 documents in one pass, 2) query the full index with 10,000 random probes and record which of four detectors finds a near-duplicate for each probe, and 3) calibrate the whole stack against synthetic injections at three perturbation strengths.

Our contributions are the following. First, we introduce capture–recapture as a measurement instrument for corpus filters, an application for which we can find no precedent in NLP, and we calibrate it on a population whose size we set ourselves, which establishes both what the estimator recovers and where it stops working. Second, we show that sampling inversion is necessary, and that omitting it understates residual duplication in FineWeb by a factor above 1,500. Third, we report a corpus-level result: an average FineWeb document has at least 68 near-duplicates elsewhere in the corpus, a bound that holds without any assumption about how duplication is distributed. The bound is on a mean degree, however, and says nothing about what share of documents carries a duplicate, which we treat throughout as unmeasured. Fourth, we read the mechanism off a positional signature rather than assuming it; the signature supports the interpretation that the survivors are predominantly cross-snapshot pairs, left by FineWeb's documented decision to deduplicate each snapshot independently after cross-snapshot deduplication had measurably hurt downstream models. Finally, we release fingerprints, capture histories and estimation code, so that any corpus that can be streamed once may be audited the same way.[1]

2 Related Work

Deduplication and its measurement. Deduplication pipelines are built from four families of method: exact substring removal (Raffel et al., 2020), document-level MinHash with locality-sensitive hashing (LSH) (Lee et al., 2022; Broder, 1997), SimHash (Charikar, 2002), and embedding-based semantic deduplication (Abbas et al., 2023). The downstream value is established: removing duplicates lowers memorization and leaves accuracy intact or better (Lee et al., 2022). Evaluation of these pipelines, however, measures yield rather than residue. Elazar et al. (2024) probe what large corpora contain, duplication included, and report rates found by exact n-gram matching, but without estimating what that matching misses. Corpus documentation studies such as Dodge et al. (2021) reconstruct a pipeline's effects in depth, and pipeline ablation studies such as Parmar et al. (2024) and Longpre et al. (2024) measure how construction choices change downstream models. None of them estimates detector recall, which is the quantity that converts a found rate into a corpus property.

Contamination as the adjacent problem. Work on benchmark contamination shares our structure, since it also asks how much of something a detector failed to see (Jacovi et al., 2023; Deng et al., 2024; Riddell et al., 2024). Chang et al. (2023) infer membership at scale from an indirect signal. That literature has converged on better detectors; the complementary move, we argue, is to estimate the recall of whatever detector is in hand.

Capture–recapture. The estimator has a long history in census undercount and epidemiological prevalence work (Fienberg, 1972; Chao, 1987), reviewed with its failure modes by Hook and Regal (1995). The closest precedent outside those fields is software inspection, where multiple imperfect inspectors serve as capture occasions to estimate undetected defects (Petersson et al., 2004). To our knowledge the method has not been applied to text corpora. Deduplicated corpora are the difficult case for this transfer. The duplicates that remain are the ones a competent detector already failed to catch, which is exactly the situation an estimator built for undetected populations addresses.

3 Two Recall Problems

Let the corpus contain M documents. Build the near-duplicate graph on them, joining two documents when their similarity exceeds the threshold, and write $K - 1$ for the degree of a uniformly chosen document, so that K counts the document to-

[1] Code and artifacts: https://github.com/garyzhang1006/residual-duplication-audit

gether with its near-duplicate neighbors. We work with degree rather than with cluster membership, and the choice is deliberate. Near-duplication at a fixed Jaccard threshold is not transitive, so the corpus does not partition into duplicate families, and any quantity defined through such a partition would depend on how one closed the relation. Degree needs no such choice. The quantity a corpus card should report is thus $\mathbb{E}[K-1]$, the mean number of near-duplicates a document has. A related quantity is $\Pr(K \geq 2)$, the share of documents with at least one. Exact-hash identity, by contrast, is a genuine equivalence relation, so exact duplicates do partition into clusters, and we use cluster language only there.

A study of the corpus through a uniform sample of $m = pM$ documents observes neither. A sampled document sees a partner only if at least one of its neighbors was also sampled, which happens with probability $1-(1-p)^{K-1}$. Writing r for the observed within-sample rate,

$$r \;=\; \mathbb{E}\big[1-(1-p)^{K-1}\big] \;\leq\; p\,\mathbb{E}[K-1]. \quad (1)$$

The inequality is Bernoulli's and it holds for every p and every K, so r/p is a lower bound on $\mathbb{E}[K-1]$ unconditionally rather than in a small-p limit. Equation 1 is the observation-recall correction, and it is model free: only the first moment enters, and nothing is assumed about the shape of the degree distribution. Documents with degree large enough to saturate the exponential contribute less than proportionally, which is what makes the direction of the bound safe.

The second correction sits inside r itself. A study measures $\hat{r}$, the rate at which its detectors flag a partner, and $\hat{r} \leq r$ whenever detection is imperfect. Capture–recapture recovers r from the overlap structure of several detectors. The two corrections compose, since the first acts on the sampling design and the second on the instrument, and neither is identifiable from the other's data.

4 Method

Corpus and fingerprints. All 14,868,862 documents of `sample-10BT` are streamed once. Per document we compute an exact hash of the first 8,000 normalized characters, a 64-permutation MinHash signature over word 5-grams, and a 64-bit SimHash over character 4-grams. Indexes are rebuilt in memory from the fingerprints, so the design needs one pass and stores no text beyond the 500-character heads used for verification. Peak memory stays under 16 GB and the full run completes in 6.3 hours on commodity CPU. Every random choice derives from fixed seeds recorded in the released code, so the run reproduces exactly.

Sampling rate. `sample-10BT` is a uniform random sample of whole documents from the FineWeb release described by Penedo et al. (2024), which draws on 96 Common Crawl snapshots running from CC-MAIN-2013-20 to CC-MAIN-2024-18. Fixing the denominator needs care, however, because the public dataset has since grown to 110 snapshots and its headline row count now covers dumps that postdate the sample. Summing the per-snapshot row counts over exactly those first 96 snapshots gives $M = 2.277 \times 10^{10}$ documents, against 2.589×10^{10} for the corpus as it now stands. We confirmed that `sample-10BT` contains no document from a post-2024-18 snapshot, so the smaller figure is the correct denominator, and it yields $p = 6.529 \times 10^{-4}$, or one document in 1,532. Using the current row count instead would inflate every corpus-level number in this paper by 13.7%. The entire inversion rests on the uniformity of the sample as well as on M, and we therefore test uniformity empirically in §5.5 rather than relying on the dataset card.

Four capture occasions. For each probe document we record which detectors find at least one other document that qualifies as a near-duplicate. Detector E is exact-hash identity. Detectors A and B split the MinHash signature into two disjoint 32-permutation families, each with its own 16-band LSH index, and each verifies candidates against its own half-signature at estimated Jaccard ≥ 0.35. Sixteen bands over 32 permutations puts two rows in a band, which places the soft LSH threshold near $(1/16)^{1/2} = 0.25$, below the verification threshold of 0.35. The band step therefore proposes candidates a little below the estimand, and the verification step does the cutting. The interval between the two is the boundary region characterized in §5.8. Because the families use disjoint random permutations, they are independent draws of the same detection process conditional on a document's true similarity to its nearest neighbor, and that conditional independence is what identifies the unobserved cell. Detector S buckets documents by the top 16 bits of the SimHash, accepts a candidate within Hamming distance 8, restricts itself to documents of at least 300 characters, and verifies every hit with a character-shingle Jaccard of at least

0.2 so that SimHash collisions cannot inflate captures. The single-block bucketing is a deliberate weakness rather than an oversight. A guaranteed Hamming-8 search would need nine blocks by the pigeonhole argument, whereas one 16-bit block finds a pair at distance d only when all d differing bits fall outside it, which happens for 8.5% of pairs at $d = 8$. Capture–recapture tolerates such a channel, since an occasion may have any recall so long as its misses are accounted for, and the consequence is the small S-only cell in Table 1 rather than bias.

Candidate caps. Both index lookups cap a bucket at 2,000 candidates. Boilerplate collides heavily in LSH, and without a cap a handful of probes would dominate the runtime of the whole pass. A capped bucket may exclude a genuine partner, but it can only remove captures and never create them, which makes every rate we report a lower bound for this reason as well as the others. The original run did not count how often a cap was reached, and we report that as an instrumentation gap rather than as a threat, since the direction of the effect is fixed. The frequency still matters for judging whether the resulting bias is negligible, however, so the released pipeline now counts cap hits per detector and writes them alongside the capture histories. Any rerun will report the number this run does not.

Estimation. The 10,000 capture histories populate a 2^4 table (Table 1). Three estimators of r are reported. The naive rate counts probes flagged by any detector. The Chapman-corrected Lincoln–Petersen estimator uses only detectors A and B, the pair whose conditional independence follows from the construction rather than from an assumption. The Chao lower-bound estimator uses singleton and doubleton capture counts and requires no model selection (Chao, 1987), following the Good–Turing argument that the frequency of once-seen units carries the information about unseen ones (Good, 1953). For completeness we also fit Poisson log-linear models over the observed cells, with all subsets of up to two pairwise interactions, and read the unobserved cell from the fitted intercept (Fienberg, 1972). Among these we select both by the Akaike information criterion (AIC) and by AIC weighting (Buckland et al., 1997). Every estimator carries a 2,000-draw bootstrap over probes that reruns the full estimation, model selection included, on each resample.

History	n	History	n
none	9,550	A+B	240
A only	72	A+B+S	82
B only	48	B+S	1
S only	3	E+A+B	1
		E+A+B+S	3

Table 1: Capture histories for 10,000 probes. E = exact hash, A/B = the two disjoint MinHash families, S = verified SimHash. The large A+B overlap and the near-symmetric A-only and B-only cells are the pattern independent channels should produce.

Injection calibration. A 6,000-document pool is reserved during streaming. We generate synthetic near-duplicates by deleting 5%, 15% or 30% of words with occasional local reordering, and run the full detection stack on 2,000 injections per strength. The recovery rates define the instrument's scope.

5 Results

5.1 What the sample shows

Direct detection flags 450 of 10,000 probes, a rate of 4.50% with bootstrap 95% CI 4.10–4.90. The Chapman estimator on the two MinHash families gives 4.58% (4.16–4.98), and Chao gives 4.81% (4.37–5.26). Table 2 collects the estimators with their intervals. The three agree to within a third of a percentage point. Such agreement indicates that the detector stack is close to saturation within its similarity scope, so the residual uncertainty about the sample is small.

The detector marginals explain where that agreement comes from. MinHash family A captures 398 probes and family B captures 375, and 326 probes are captured by both. Exact hashing captures 4 and verified SimHash captures 89, of which only 3 are captured by SimHash alone. The design therefore has two strong, symmetric occasions and two weak ones, which is enough for Chao and Lincoln–Petersen, but not enough for a four-way log-linear fit.

5.2 Validation on a known population

Before trusting the estimator on a population we cannot see, we ran it on one we constructed. Each injection batch contains exactly 2,000 planted near-duplicates, so the true population is known and the estimators can be scored against it (Table 3).

At 5% word deletion the detectors find 1,990 of 2,000 and nothing is left to estimate, so every method returns the same answer. The informative

Estimator	Rate	Bootstrap 95% CI	Corpus $\mathbb{E}[K-1] \geq$
Detected (naive)	4.50%	4.10–4.90	68.9
Lincoln–Petersen	4.58%	4.16–4.98	70.1
Chao (primary)	4.81%	4.37–5.26	73.7
Chao (bias-corr.)	4.81%	4.37–5.26	73.7
Chao, dep.-adjusted	5.40%	—	82.7

Table 2: Within-sample near-duplicate rate by estimator, with 2,000-draw bootstraps that rerun the full estimation on each resample, and the corpus-level bound each rate implies through Equation 1. The upper block is what we report. The final row divides the Chao rate by the 89.2% recovery it achieved on the planted population of §5.2, which is our best available handle on detector dependence and heterogeneity; §5.7 explains why we treat it as a sensitivity rather than as the headline. Two log-linear variants are unstable under resampling and appear in Appendix B.

Deletion	Truth	Detected	Lincoln–P.	Chao
5%	2,000	1,990 (99.5%)	1,990 (99.5%)	1,991 (99.5%)
15%	2,000	890 (44.5%)	1,332 (66.6%)	1,784 (**89.2%**)
30%	2,000	6 (0.3%)	15 (0.8%)	21 (1.1%)

Table 3: Estimators scored against a known population of 2,000 planted near-duplicates per batch. At 15% deletion, where detection misses more than half, Chao recovers 89.2% of the truth from a 44.5% detected rate while remaining a lower bound.

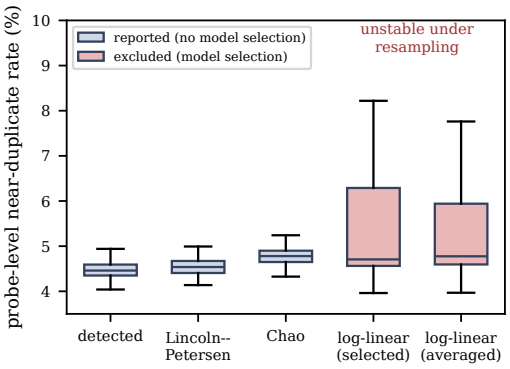


Figure 1: Bootstrap distributions over 2,000 probe resamples. The three estimators on the left are stable and mutually consistent. Both log-linear routes, including AIC model averaging, inherit the instability of sparse multi-detector cells and are excluded from the headline.

case is 15% deletion, where direct detection finds 890 and misses more than half of the planted population. Chao recovers 1,784, which is 89.2% of the truth, and Lincoln–Petersen recovers 1,332, or 66.6%. Both stay below the true value, as the theory requires of a lower bound under heterogeneous catchability, and Chao doubles the naive count in the right direction. At 30% deletion the instrument is blind, only six injections are found at all, and no estimator can repair a design that captured almost nothing.

One table thus establishes both the credibility of the corpus estimates and their scope. Where detection is partial, Chao closes most of the gap; where detection has collapsed, it correctly cannot. The scope of that claim, however, is the perturbation family we planted. Real web duplicates also arise through templating, boilerplate edits, partial copying, truncation and markup change, so the 89.2% figure calibrates the estimator under word deletion with local reordering rather than measuring its recall on FineWeb's own duplication process. What the table licenses is the direction and the rough magnitude of the correction, not a transferable recall figure.

5.3 Which estimator to trust

The log-linear route fails this test, and Figure 1 shows why we exclude it. AIC selection gives 4.64% at the point estimate, but a bootstrap interval of 4.26–8.85, because five cells hold fewer than five probes and the selected model changes under resampling. AIC weighting, the standard remedy for selection instability, does not help; Appendix B gives both variants with their intervals and the across-model spread. Reporting a log-linear headline here would be reporting an artifact of sparse overlap cells. We therefore take Chao as the primary estimator, since it needs no model selection and is a lower bound under heterogeneous catchability, and Lincoln–Petersen as a check whose independence assumption can at least be derived from the construction. That derivation is conditional on a document's latent similarity to its nearest neighbor. Marginal dependence across documents survives it, so both estimators should be read as conservative lower bounds rather than as unbiased estimates.

5.4 The survivors are cross-snapshot

Scanning the full corpus, 51,972 documents share their first 8,000 normalized characters byte-for-byte with an earlier document, and 103,552 documents, or 0.70% of the corpus, sit in an exact-duplicate cluster of size two or more. The cluster spectrum is almost pure pairing: 51,279 clusters of size two,

281 of size three, and a tail reaching a single cluster of 36.

Where those pairs sit in the stream carries information about their origin, but how much it carries depends on how the stream is laid out. We measured the layout directly, by reading the snapshot label of 40,000 documents in runs of 100 at 400 positions spaced evenly through `sample-10BT`. The label changes 266 times in 39,600 adjacent pairs, so one snapshot holds the stream for a mean run of 149 documents, and 139 of the 400 runs stay inside a single snapshot. Snapshots interleave at that fine grain, and they do not run the length of the corpus. The sampled documents of a given snapshot span a median 6.3% of the stream and at most 8.8%, and the regions appear in an order unrelated to crawl date. Position is therefore a proxy for snapshot at the scale of a few percent of the stream, not at the scale of a block of twenty. A within-snapshot duplicate pair would fall inside one of those regions, contributing separations under 9%, so failed within-snapshot deduplication would pile mass into the left tail of the histogram. Figure 2 shows no such pile. Over the 51,279 clusters of size exactly two, the observed mean separation is 34.4% of the corpus span against 33.3% expected under uniform document sampling, and the observed median is 30.3% against 29.3% expected. Only 1.4% of pairs fall within 1% of the span.

Two readings follow. First, the histogram supports the interpretation that the surviving exact duplicates are predominantly cross-snapshot residue, with the 3% excess over the uniform curve reading as depletion of the same-snapshot diagonal by a deduplication stage that worked. This matches FineWeb's documented design, in which each snapshot was deduplicated independently after global deduplication had been tried and rejected for harming downstream performance. Second, and needed for everything in §5.6, the agreement with the uniform curve is evidence that `sample-10BT` samples documents rather than files or shards. Shard-level sampling would concentrate pairs and produce a visibly different histogram.

The measured region width also bounds how much within-snapshot residue the histogram tolerates. Mixing a share f of within-snapshot pairs, each uniform inside a region of width 0.09 and so averaging 0.03, with cross-snapshot pairs averaging the uniform 0.333, gives an overall mean of $0.333 - 0.30f$. The observed mean of 0.344 lies above the uniform value rather than below it, so the model returns no positive f at all. The margin is wide: over 51,279 pairs the standard error of the mean separation is 0.0010, and even a one percent within-snapshot share would sit thirteen standard errors below what we see.

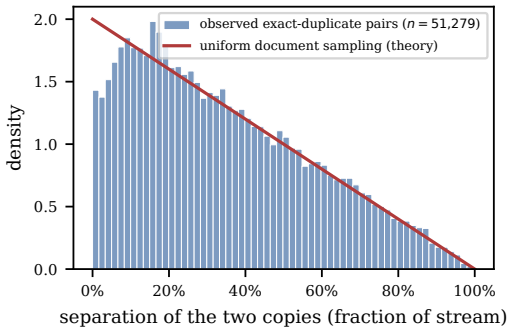


Figure 2: Separation between the two copies of each exact-duplicate pair, as a fraction of the corpus stream. The observed distribution tracks the triangular density expected under uniform document sampling, with a slight shift away from zero. One snapshot spans at most 8.8% of the stream (§5.4), so surviving within-snapshot pairs would show as mass in the left tail, and the tail is if anything thinner than uniform.

Position nevertheless remains a proxy, and a direct measurement exists. FineWeb rows carry a `dump` field naming the snapshot each document came from, so storing one label per fingerprint would replace the separation histogram with the fraction of exact-duplicate pairs whose two members disagree on that label. The run reported here did not store it, but the released pipeline now does, so the next application of this audit can report the direct fraction. We prefer that number to ours, and we report the proxy as a proxy.

5.5 Is the sample uniform?

The inversion multiplies by $1/p$, a factor above 1,500, so uniformity of the sample carries more weight here than a design assumption usually does. The separation histogram of §5.4 tests it positionally, which leaves open any non-uniformity by document length, domain or quality that keeps the separations intact. The release publishes the snapshot label and the token count for every row, however, and the 40,000 labelled documents described above supply both.

Composition by snapshot is the sharper test, since crawl date carries much of what varies across the corpus. We group the 96 snapshots by crawl year and compare each year's share of the sample

against its share of the release. Ten of the twelve years agree within a 95% interval from a bootstrap that resamples whole runs rather than documents, and neither of the two that miss is off by as much as a factor of two (Appendix E). Length agrees more closely: the sampled documents average 700 GPT-2 tokens against a release mean of 659, so the sample runs about 6% long, which we treat as an upper estimate of the length skew because the published token total is rounded. Neither check covers domain or quality score, and 400 runs resolve a single snapshot only coarsely. What they do rule out is a sampler distorted by a factor rather than by a tenth, the regime that would matter against a correction of 1,500.

5.6 From the sample to the corpus

Applying Equation 1 with $p = 6.529 \times 10^{-4}$ converts the within-sample rates into corpus quantities: the naive rate gives $\mathbb{E}[K-1] \geq 68.9$, Lincoln–Petersen gives ≥ 70.1, and Chao gives ≥ 73.7. An average FineWeb document, then, has at least 68 near-duplicates elsewhere in the corpus. That should be read as a lower bound on a mean degree and not as a statement about a typical document, since a mean over a heavy-tailed degree distribution can be carried by a minority of documents in large clusters. The same inversion applied to exact duplicates gives $\mathbb{E}[K-1] \geq 10.7$, so the average document also has at least ten byte-identical copies of its first 8,000 characters spread across the corpus.

The mean does not by itself pin down how many documents are affected. Since $\Pr(K \geq 2) \cdot \mathbb{E}[K-1 \mid K \geq 2] = \mathbb{E}[K-1]$, converting a mean of 68 into a prevalence would require the mean degree among documents that have any neighbor at all, and our binary capture design does not measure that quantity. The resulting range is wide. A conditional mean of 95, which one surviving copy per snapshot across 96 snapshots would give, implies a prevalence of 72.5%, while conditional means of 300 and 1,000 imply 23.0% and 6.9%. The whole range is compatible with our data, so we report $\mathbb{E}[K-1]$ as the result and treat prevalence as unmeasured.

Two pieces of evidence favor the lower end of that range without settling it. The exact-duplicate spectrum puts at most 5% of its own first moment in clusters with three or more co-sampled members, so the observable heavy tail is not what carries the mean. Independently, the corpus authors report that deduplicating across snapshots removed more than 90% of the base filtered data in the oldest snapshots. Community measurements on FineWeb-Edu find that roughly 75% of documents share text with another document.[2] A high prevalence with moderate degree fits both observations, whereas a low prevalence with degree in the thousands fits neither. Recording neighbor counts rather than binary capture would settle it in one additional pass.

5.7 How far the inversion can be pushed

Three quantities control the corpus-level numbers. The sampling rate enters linearly, so an error of 10% in the corpus size moves $\mathbb{E}[K-1]$ by 10% in the same direction. It is also the one sensitivity we can check rather than assume, and the check matters: summing per-snapshot row counts over the 96 snapshots the sample was drawn from, rather than over the corpus as it stands today, changes the answer by 13.7%.

Saturation makes Equation 1 a bound, and the size of the gap depends on where the degree mass sits. At the per-snapshot ceiling of 96 the linear approximation overstates the detection probability by 3.0%, so the bound is nearly tight for exactly the degrees the mechanism predicts, while the gap widens to 26.5% at degree 1,000 and to 97.2% at degree 55,000, the largest cluster our spectrum implies. The true mean therefore sits above our bound by an amount that grows with the weight of the tail.

Detector dependence is the third quantity, and the planted populations of §5.2 let us put a number on it rather than only a direction. At 15% deletion, where the same four detectors face the same latent-similarity correlation they face on the corpus, Chao recovers 89.2% of a known population and Lincoln–Petersen recovers 66.6%. Read the Chao shortfall as an estimate of the combined dependence and heterogeneity bias, and divide through by it. That moves 4.81% to 5.40% and $\mathbb{E}[K-1]$ from 73.7 to 82.7, which Table 2 carries as its final row. We do not adopt it as the headline, however, since a planted population of single twins is neither the mixture of degrees the corpus presents nor a sample of more than one perturbation family. It does bound the discussion: the direction is up, and the

[2] Reported in the public discussion of the dataset repository (`fineweb-edu`, discussion 7 on the HuggingFace hub), where a dataset author confirms that the release deduplicates each snapshot individually.

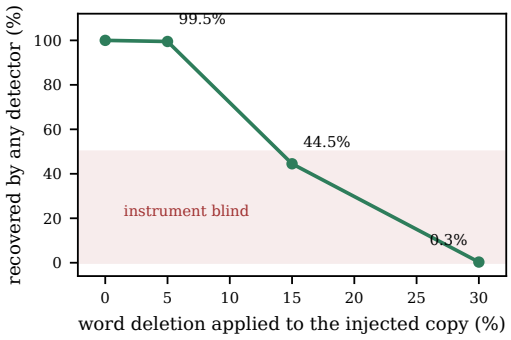


Figure 3: Recovery of synthetic near-duplicates as a function of perturbation strength, 2,000 injections per point. Sensitivity collapses between 15% and 30% word deletion, which fixes the similarity scope of every estimate we report.

magnitude where we can check it is 11%, not a factor.

The tail is real. Fitting a single cluster size to the exact-duplicate spectrum through the S_2/S_1 ratio gives $\kappa = 11.6$. That model then underpredicts the higher cells badly, by a factor of 2.6 at three co-sampled members and 59 at four, and it puts clusters of five and six at effectively zero probability (Appendix D). Cluster sizes are heavy-tailed. That is what web duplication should look like, and it is what makes $\mathbb{E}[K-1] \geq 68$ conservative.

5.8 What the instrument can see

Injection recovery is 99.5% at 5% word deletion, 44.5% at 15%, and 0.3% at 30% (Figure 3). The estimand is therefore near-duplication at Jaccard similarity of roughly 0.35 and above, and the instrument says nothing about heavier paraphrase. Every claim above is conservative as a result, since duplicates too heavily rewritten for any of our detectors are counted as absent.

Recovery is only half of a calibration, and the other half matters here because capture fires on a noisy statistic estimated from 32 permutations. Appendix C works through the false-positive side and finds it negligible: threshold noise on unrelated documents does not manufacture captures. What it does instead is blur the boundary between roughly 0.25 and 0.35 in true similarity, which moves where the estimand sits by a little.

6 Discussion

The result changes what a deduplication report can be read to mean. FineWeb's pipeline is not broken, and the duplicates that remain follow from a deliberate choice: the authors tried deduplicating across snapshots, found that it removed high-quality text and hurt downstream performance, and switched to per-snapshot deduplication. The corpus card, however, does not state the size of the residue that choice leaves. A practitioner who reads that FineWeb is deduplicated has no way to anticipate an average of 68 near-duplicates per document.

The two corrections differ enormously in size. Detector recall moves the number by seven percent relative, from 4.50% to 4.81%, for which four detectors and a 2^4 table are the price. Observation recall, by contrast, moves it by a factor above 1,500.

That asymmetry suggests a reporting convention that would cost almost nothing. A release reporting residual duplication should give four things: the rate its detectors found, the fraction p of the corpus they ran on, the exact release the denominator M came from, and the similarity threshold at which a pair counts as duplicate. With those four numbers a reader can apply Equation 1 without access to the corpus. Without them a published rate cannot be compared across corpora, or across versions of one corpus. Pinning the release matters: FineWeb grew from 96 to 110 snapshots between the paper and the dataset card, and taking the denominator from the card would have inflated every corpus number here by 13.7%. A corpus that ships a residual estimate this way makes a falsifiable claim about its own cleanliness, which is the property none of the releases we surveyed has.

For a practitioner the finding has immediate uses. Anyone auditing duplication in a subsample now has a reason to expect their own rate to understate the corpus, and the sampling rate they already know converts one into the other. Anyone choosing between per-snapshot and cross-snapshot deduplication can weigh the downstream loss the FineWeb authors reported against the residue their alternative leaves, where previously only the first had a number. And our released fingerprints identify which `sample-10BT` documents have a surviving near-duplicate, so memorization work can compare the two halves without recomputing anything.

The design should generalize, although one corpus and one deduplication strategy leave the broader claim resting on a single case. It needs one streaming pass, no stored text and no GPU. Running it across RefinedWeb and Dolma, which both deduplicated globally, would turn residual duplication into a property of pipeline design.

Limitations

The corpus-level claims in §5.6 rest on the sampling inversion, and the inversion rests on two conditions. The first is uniformity. `sample-10BT` must be a uniform sample of documents, which we test two ways and find supported both times. The first test, in §5.4, is positional. It would not detect a non-uniformity in sampling probability by document length, domain or quality that leaves the separation distribution intact. This matters more than a usual design assumption, because p multiplies the headline by a factor above 1,500. A sampler favoring longer or higher-scoring documents, even by a modest margin, would move the bound further than any other quantity reported here. The stronger test compares document covariates between the sample and the full release, and §5.5 reports what that comparison shows on the snapshot label and the token count. Those are the two covariates the release publishes for every row. Domain and quality score are not covered, and they remain untested. The second condition is the corpus size, since p enters every corpus-level number linearly. M is computed by summing published per-snapshot row counts over the 96 snapshots the sample was drawn from, which is verifiable, but not something our own data can confirm. Taking the current headline row count of the still-growing dataset instead would inflate every corpus number here by 13.7%.

Prevalence is the claim we cannot make. As §5.6 sets out, $\Pr(K \geq 2)$ follows from $\mathbb{E}[K-1]$ only through the conditional mean degree, which our binary capture design never observes, and the plausible range spans 7% to 73%. Readers should take $\mathbb{E}[K-1] \geq 68$ as the result of this paper. It should not be converted into a share of the corpus. An earlier version of this work reported such a share, derived from a ceiling of one surviving copy per snapshot. Our own spectrum contains a cluster of 36 co-sampled members, implying roughly 55,000 copies corpus-wide, which violates that ceiling and invalidates the derivation.

The capture–recapture side has its own limits. Detectors A and B are described here as independent capture occasions, but that holds only conditionally, given a document's true similarity to its nearest neighbor. Across documents, all content-based detectors correlate through that latent similarity. The marginal dependence is therefore positive, and both Chao and Lincoln–Petersen are biased downward, which compounds with the lower-bound direction of Equation 1. Detector S contributes little independent information, with only three probes captured by SimHash alone, so the effective design has closer to two occasions than four.

Our specificity check in §5.8 runs on the 6,000-document held-out pool rather than on the corpus index. A probe against the full index draws candidates that already share an LSH band, so their similarity distribution is shifted upward relative to a random pool pair. The pool measurement bounds noise on unrelated text. It does not characterize the boundary region between roughly 0.25 and 0.35, which is where the threshold does its real work. Repeating the check against the full index, which our released fingerprints permit, would close that gap.

The exact-hash detector keys on the first 8,000 normalized characters, which creates a threat to §5.4 beyond simple miscounting. Templated sites can share a long prefix across unrelated articles. Such pairs would also scatter across snapshots and reproduce the positional signature we read as evidence of cross-snapshot survival. Among the 6,000 held-out pool documents, prefix agreement decays fast: three pairs share the first 100 characters, two share 200, and none share 500 or more out of nearly 18 million pairs. The evidence is real and weak at once, however, because a random sample of a corpus separates template siblings just as efficiently as it separates duplicates, so the pool cannot contain many of either. Settling the question needs a check against the full index on document text, and our released fingerprints do not carry it.

The scope, finally, is one corpus at one time, and the deduplication design studied here is a single one. The general claim that deduplication reports understate residual duplication is therefore supported by one case rather than by a comparison across pipelines. We measure near-duplication at Jaccard 0.35 and above and say nothing about paraphrase, which a semantic detector added as a fifth occasion could reach, and which matters more each year as generated text enters web crawls. Duplicates are planted by word deletion with local reordering, so the 89.2% recovery describes the estimator under that family and not under templating, boilerplate edits, truncation or markup change. The calibration also has a resolution gap. Recovery falls from 44.5% at 15% deletion to 0.3% at 30%, and no point in between was measured, so

the sensitivity boundary that scopes every claim in this paper is located only to within that interval. Residual duplication is not connected here to any downstream training outcome either. The paper measures a corpus property without establishing what it costs.

Ethics Statement

This work analyzes a public dataset and releases no text beyond what FineWeb already publishes. The released artifacts are hashes, capture histories and code. Identifying which documents are duplicated could in principle assist reconstruction of corpus membership, but the fingerprints are one-way and the underlying corpus is already open.

The quantity we measure has downstream stakes beyond documentation practice. Repeated text drives verbatim memorization, the mechanism by which personal information and copyrighted passages reappear in model output (Carlini et al., 2023; Lee et al., 2022). It is also the route by which evaluation material leaks into training sets and inflates reported capability (Jacovi et al., 2023). A residue the corpus card does not mention is a risk that no model builder can price. Measuring it, as this paper does, is the step that makes the risk visible. Reducing it is a separate decision with its own cost, since the FineWeb authors found that removing cross-snapshot duplication hurt the models they trained.

Acknowledgments

I thank the four WaC-13 reviewers, whose reports prompted the sampling-uniformity analysis in §5.5, the dependence-adjusted row of Table 2, and a good deal of tightening elsewhere.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. 2023. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Multimodal Representation Learning*.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29.

Steven T. Buckland, Kenneth P. Burnham, and Nicole H. Augustin. 1997. Model selection: An integral part of inference. *Biometrics*, 53(2):603–618.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327.

Anne Chao. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791.

Moses S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 380–388.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719.

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.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Stephen E. Fienberg. 1972. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59(3):591–603.

Irving J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Ernest B. Hook and Ronald R. Regal. 1995. Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews*, 17(2):243–264.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8424–8445.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, and toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Bo Liu, Aastha Jhunjhunwala, Zhilin Wang, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Data, data everywhere: A guide for pretraining dataset construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10695.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. The fineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only. *Advances in Neural Information Processing Systems*, 36.

Håkan Petersson, Thomas Thelin, Per Runeson, and Claes Wohlin. 2004. Capture-recapture in software inspections after 10 years research: Theory, evaluation and application. *Journal of Systems and Software*, 72(2):249–264.

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.

Martin Riddell, Ansong Ni, and Arman Cohan. 2024. Quantifying contamination in evaluating code generation capabilities of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14116–14137.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788.

A Reproducibility

Stage A streams `sample-10BT` once and writes fingerprints; stages B through D read only those fingerprints. The full run takes 6.3 hours on a single commodity CPU with peak resident memory under 16 GB, and it needs no GPU and no network access after the corpus is available locally. Two seeds fix every random choice, `20260811` for the corpus pass and `20260812` for estimation. The first governs the MinHash permutation family, the 10,000-probe sample, the 6,000-document injection reservoir, and the perturbations applied to injected copies; the second governs the 2,000-draw bootstrap. `PYTHONHASHSEED` is pinned to 0 by a re-exec at process start, so string hashing is stable across machines, which the stage-resume logic depends on. Re-running `estimate.py` against the released capture histories reproduces every number in Tables 1, 2, 3 and 4 without touching the corpus, and `figures.py` regenerates the figures from the same artifacts. The repository holds the pipeline, the capture histories, the injection results, the bootstrap draws, and the estimation code.

B Log-linear estimators

Table 4 gives the two log-linear variants that §5.3 excludes. We fit Poisson log-linear models over the observed cells of Table 1 with all subsets of up to two pairwise interactions, reading the unobserved cell from the fitted intercept (Fienberg, 1972), and we select among them both by AIC and by AIC weighting (Buckland et al., 1997). Both point estimates sit inside the range the stable estimators cover, and both bootstrap intervals run to roughly 8.7%, nearly double the point estimate. The cause is sparsity rather than a defect of the models: five cells of the 2^4 table hold fewer than five probes, so a resample can move the selected model between specifications whose implied unobserved cells differ by an order of magnitude. AIC weighting averages over that instability instead of removing it, and the across-model standard deviation of the unobserved cell is 51.5 against a point estimate of 38.8. A design with three strong occasions rather than two would support this family of estimators; ours does not.

C Specificity of the capture rule

Recovery is only half of a calibration, and the other half matters here because capture fires on a noisy statistic. A 32-permutation half-signature

Estimator	Rate	95% CI	$\mathbb{E}[K-1]$
Log-linear, AIC-sel.	4.64%	4.26–8.85	71.1
Log-linear, AIC-avg.	4.89%	4.28–8.64	74.9

Table 4: Log-linear variants, excluded from the headline. Bootstrap intervals reach nearly twice the point estimate because model selection is unstable over sparse overlap cells.

estimates Jaccard with a standard error near 0.084, so a document with many candidates gets many chances to clear 0.35 by chance. We measured the resulting false-positive rate two ways. Analytically, capture requires at least 12 of 32 minima to agree, which for a pair with true Jaccard J has probability $\Pr(\mathrm{Bin}(32, J) \geq 12)$: exactly zero at $J = 0$, 2.1×10^{-8} at $J = 0.05$, and 3.3×10^{-5} at $J = 0.10$. Empirically, we compared all 17,997,000 pairs among the 6,000 held-out pool documents under both MinHash families, 36 million comparisons in total, and the rule captured none of them. Threshold noise on unrelated documents does not manufacture captures. It blurs the boundary between roughly 0.25 and 0.35 in true similarity, which shifts where the estimand sits by a little and leaves unrelated text outside it.

D Exact-duplicate cluster spectrum

Table 5 gives the spectrum behind the heavy-tail claim of §5.7.

Co-sampled members j	Observed clusters	Single-size prediction	Ratio
1	14,765,310	14,765,310	1.00
2	51,279	51,279	1.00
3	281	108	2.6
4	9	0.152	59
5	4	1.5×10^{-4}	2.6×10^{4}
6	2	1.1×10^{-7}	1.8×10^{7}

Table 5: Exact-duplicate cluster spectrum in the sample against a single-cluster-size model calibrated on S_2/S_1 ($\kappa = 11.6$). The model is fitted through the first two cells and fails from the third onward, which identifies a heavy tail in cluster size.

E Composition of the sample by crawl year

Table 6 compares the share of `sample-10BT` contributed by each crawl year against the share the 96-snapshot release contributes, using the 40,000 documents whose snapshot label we read at 400 positions through the stream. Intervals come from a 2,000-draw bootstrap that resamples the 400 runs rather than the 40,000 documents, since documents inside a run share a snapshot and are not independent. The comparison is the covariate check §5.5 describes.

The two years whose release share falls outside the bootstrap interval are 2021, where the sample holds 7.1% against 10.2% in the release, and 2024, where it holds 1.2% against 2.3%. The larger gap falls in the year contributing 2% of the corpus. Over all 14,868,862 documents the sampled mean of 700 GPT-2 tokens implies 1.04×10^{10} tokens, within 4% of the ten billion the release name promises; the release reports roughly 1.5×10^{13} tokens over 2.277×10^{10} documents.

Year	Release	Sample	95% CI	Ratio
2013	1.91%	1.78%	0.8–3.0	0.93
2014	7.36%	7.43%	5.3–9.7	1.01
2015	8.25%	8.70%	6.8–11.0	1.05
2016	7.45%	8.19%	6.1–10.4	1.10
2017	14.46%	15.86%	12.9–19.1	1.10
2018	13.04%	12.13%	9.4–15.0	0.93
2019	11.63%	11.23%	8.6–13.9	0.97
2020	9.16%	10.32%	7.7–12.9	1.13
2021	10.20%	7.08%	5.0–9.3	**0.69**
2022	7.42%	9.37%	6.8–12.1	1.26
2023	6.82%	6.74%	4.6–9.1	0.99
2024	2.30%	1.17%	0.4–2.1	**0.51**

Table 6: Share of documents by crawl year, in the release and in `sample-10BT`. Bold marks the two years whose release share falls outside the bootstrap interval.

Using search engines to trace metalinguistic discussions of the noun *female* in digital discourse

Marie Flesch
Université de Lorraine / Nancy, France
marie.flesch@univ-lorraine.fr

Abstract

This contribution uses the web as corpus to investigate metalinguistic discourses of the noun "female" between 2001 and 2019. It presents the construction of a diachronic dataset using search engines and compares results provided by Google, DuckDuckGo and Yandex, and, for Google, searches made with four different combinations of parameters. It then uses the resulting corpus to create a repertoire of arguments and ideologies used by online critics of "female". Results show that, even though Google generated the most relevant results by far, supplementing it with other search engines was beneficial. They also suggest that, by 2015, the metalinguistic discourse of "female" had mostly solidified, with no new arguments emerging in the corpus, and that the debates are a response to new (at the time) and broader social tensions, specifically the rise of masculinist discourses in online spaces.

1 Introduction

Since the 1970s, feminist scholars have been uncovering the asymmetries in references to women and men (for example, the tendency to refer to men as "men" and to women as "girls", Lakoff, 1975; Sigley and Holmes, 2002), and the process of semantic pejoration undergone by many terms referring to women, seen in pairs like "master and mistress" (Schulz, 1990). Metalinguistic debates about terms for women long predate second wave feminism, however, and have unfolded in dictionaries, style books, and usage guides. One such debate revolves around the noun "female", which fell out of favour in the 19th century, coming to be seen as "vulgar", "contemptuous" (Hall, 1917) and "impolite" (Fowler, 1926, p. 175). Yet, when second wave feminists began reclaiming the word "woman", their efforts primarily targeted the words "lady" and "girl" rather than "female"; for example, Robin Lakoff (1975) does not discuss "female" and even uses it as a noun to refer to women. Many academic works about language and gender, including some discussing gender-inclusive guidelines, continued to use the noun in reference to women in the first decades of the 21st century (for example, Hellinger, 2010; Mills and Mullany, 2011; Baker, 2014 — as well as my own work, for ex. Flesch and Bailly, 2020). Recently, however, the debate around "female" seems to have picked up, as evidenced by newer editions of some style guides, such as the 2022-2024 edition of *The Associated Press Stylebook* and the 2024 edition of *The Chicago Manual of Style*, which now note the negative connotations of the nouns "female" and "male" and recommend to not use them in reference to humans (Froke et al., 2022; The University of Chicago Press Editorial Staff, 2024).

In attempting to retrace the developments of the renewed discussion about the noun "female", I found few traces of it in academic works, in feminist theory and beyond (for example in Salleh, 1992; Frye, 2005; Peters et al., 2023; Montell, 2024). My hypothesis is that the debate mostly re-emerged from "below", in non-academic discussions held on the internet, precisely because the noun "female" has been repoliticized online through its use by incels and other men rights activist movements, and more broadly in reaction to the development of a "networked misogyny" (Banet-Weiser and Miltner, 2016). A notable example of these grassroots efforts is the Reddit forum r/MenAndFemales, which is dedicated to calling out the use of the noun and frequently features incel content (JenJens, 2026).

This is why, in order to retrace the trajectory of the "female as a noun" debate in the 21st century, I turned to the web as corpus. Three search engines (Google Search, Yandex, and DuckDuckGo) were used to search for the expression "female as a noun" and to build a diachronic corpus of metalinguistic discussions in blogs, discussion forums,

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 182–191
October 29, 2026 ©2026 Association for Computational Linguistics

social media platforms, media articles, and comments sections. Using search engines to create such a corpus comes with obvious limitations, notably the lack of reproducibility, the instability and inconsistency of search results, and the opacity of the algorithms used (Kehoe, 2020). However, it also has advantages: it is more lightweight than using the massive Common Crawl or Reddit archives, does not require knowledge of programming languages, and provides access to discussions in their context, allowing for more situated linguistic analyses. Moreover, it provides access to content that predates Common Crawl archives, which (as of the time of this writing) are accessible only for the year 2013 and onwards[1]. While the corpus makes no claim to exhaustivity or representativity, it constitutes a valuable roadmap of the recent developments of the "female vs. woman" debate.

The contribution this paper makes is twofold:

1. It presents a comparison of the results of searches using the exact expression "female as a noun" provided by three search engines (Google, Duckduckgo, Yandex) and, for Google, using combinations of different search parameters: with and without "omitted" results; global search (all times) and searches in custom ranges of times. The goal is to establish whether, in order to get as many quality results as possible for the study at hand (specifically for older web data), it is beneficial to (a) complement Google with other search engines and (b) combine different types of Google searches. This could be useful to researchers who want to conduct diachronic studies of linguistic or sociocultural phenomena using the web as corpus.

2. It investigates metalinguistic discourse in the resulting corpus (available on OSF[2]), looking at the ways internet users frame "female" as a noun: it uncovers the repertoire of arguments used by critics and the social categories associated with the usage of the term. Results suggest that, by 2015, the metalinguistic discourse had already solidified. They also reveal that it is not only a response to evolving norms in language use, but also to broader social tensions, particularly the rise of the manosphere.

[1]https://data.commoncrawl.org/crawl-data/CC-MAIN-2012/index.html

[2]https://osf.io/hq62x/overview?view_only=ec480386076144d48e21147b4d10b44b

2 Related work

2.1 Search engines

Research has shown that results provided by various search engines can differ significantly. Because of its overwhelming market dominance, Google usually outperforms other search engines in terms of relevance, ranking and coverage of results (Hajian Hoseinabadi and CheshmehSohrabi, 2024). However, search engines are not simply entry points to the internet. Because it is also a content and platform provider, with services like Google Books, YouTube or Google Maps, Google tends to give priority to its own services and collections in its search results (Lewandowski, 2023). It has also been shown that, when looking specifically for Reddit results, Google disproportionately excludes discussions related to politics, pornography, bots, and cryptocurrency, acting as a "digital gatekeeper" that determines what information is accessible and shaping online discourse (Poudel et al., 2025).

The shortcomings of the "web as corpus" approach, which emerged in the late 1990s, have been pointed out by corpus linguists. Because the actual size of the internet is unknown, it cannot be considered a "corpus" and cannot be used to create representative samples (Sinclair, 2005). The other issue, when using search engines, is that of reproducibility: it is made impossible by the frequent changes in algorithms and by the "bubble effect", i.e. the fact that search engines tailor their results based on derived users' preferences (Ćurković and Košec, 2018). Difficulties in using search engines to create corpora is compounded by the decrease in options offered by search engines, which makes them less and less suitable for the extraction of linguistic data (Kehoe, 2020). The shift toward AI-driven search could further reduce the reliability of search engines (Memon and West, 2024).

These biases have serious implications for any research that uses solely search engines as a corpus building tool. There are alternatives, such as using Common Crawl web archives. However, using them is problematic: it requires access to big data infrastructure and skills, documents are not preprocessed, and the archives contain many near duplicates of pages (Panchenko et al., 2018). For the present study, which aims to retrace the developments of a metalinguistic discourse as far back in time as possible, the fact that Common Crawl only provides content crawled since 2013 constitutes an additional limitation.

2.2 The "female" as a noun debate

Metalinguistic debates around "female" have been documented by John Lesslie Hall (1917) and Dennis E. Baron (1986). They are largely restricted to the noun, while the adjective "seems to have no enemies" (Hall, 1917, p. 89). In use as a synonym for "woman" since the 14th century, "female" began to meet some opposition by writers in the 18th century, and then came under attack in the 19th century because of its negative connotation. The writer Sarah Josepha Hale, who carried the first campaign against the noun in the American magazine *Godey's Lady's Book*, noted that it "reduce[s] [women] to the level of animality" (LC, 2018). Other arguments have been put forward in dictionaries and style guides: it was described as "impolite" Fowler (1926, p. 175), "contemptuous" (Evans and Evans, 1957, p. 560), "clinical" (Miller and Swift, 1980, p. 76), "problematic" (American Psychological Association, 1995, p. 55), and retaining a "sense of non-humanness" (Doyle, 1995, p. 31). However, in reclaiming the word "woman", second wave feminists mostly focused on "lady" and "girl" (cf Lakoff, 1975). The debate has recently been reignited, with William Safire (2007) writing in *The New York Times* that "feminists everywhere have begun to turn on the word female". In the same article, Deborah Tannen notes that many feminists avoid the term because it "connotes a biological category". Amanda Montell also argues that the woman vs. female debate is fuelled by the "gender versus sex question" (2024), while Deborah Cameron turned to the British National Corpus to show that "female" as a noun is often used by men to make disparaging comments on women (Cameron, 2016). Recent corpus linguistics studies suggest that the noun is undergoing a process of pejoration (Robinson, 2018; Ferley, 2022; Joensuu, 2023).

However, it seems that most of the metalinguistic discourse about "female" is not taking place in academic spaces, but in digital discourse, in forums, blogs, comment sections, and social media. For example, it has been the main focus of the subreddit r/MenAndFemales since its creation in 2020. To the best of my knowledge, no study has investigated systematically the recent developments of the "female" debate: this contribution aims to address this gap, by using the web as corpus.

3 Search engines comparison

3.1 The search expression

The goal was to use an expression that would most likely provide results containing metalanguage, and not return too many false positives (general discussions about language or just simple usage of the term). From my ongoing work about the topic, "female as a noun" emerged as a phrase that is frequently used in discussions about the word. For example, it is the title of one chapter in Hall (1917)'s English usage guide; and it features in the first rule of the r/menandfemales subreddit ("'Female' as an adjective is okay. 'Female' as a noun is not"[3]). The quotation marks search operator was used in order to force exact-match searches.

3.2 The search engines

Search engines were selected on the basis of their popularity. According to statcounter (2026), the five most used search engines are Google Search with a market share of 91.31%, Microsoft Bing (market share: 4.47%), Yahoo! Search (market share: 1.24%), Yandex Search (market share: 1%) and DuckDuckGo (market share: 0.65%). However, Bing and Yahoo! had to be discarded because initial searches showed that using the quotation marks search operator did not work, returning results that contained both "female" and "noun" but not necessarily "female as a noun". The fact that both search engines behave in the same way is not surprising, seeing that Yahoo! Search uses Microsoft Bing to generate results (BBC News, 2009). The way searches were performed varied depending on the affordances of each search engine. Because the Russian search engine Yandex Search does not allow for time-specific searches, a simple global search was conducted. For DuckDuckGo, a search engine created by an independent software company that focuses on internet privacy, custom time ranges were defined: 2000-2010, 2011-2015, 2016-2020, 2021-2022, and 2022-2026. Google also offers a date range parameter; and, depending on the results of a search, it provides the option to display "omitted results", i.e., results that are not considered the most relevant. Given Google's dominance as a search engine and the fact that, in exploratory searches, it was clear that it returned a lot more results that Yandex Search and DuckDuckGo, I decided to compare results provided by

[3] https://www.reddit.com/r/MenAndFemales/

different combinations of parameters: (1) global results (without setting a date range), only showing the default results; (2) global results (without setting a date range) including the omitted results; (3) results obtained using a custom date range, only with the default results; and (4) results obtained using a custom date range, including omitted results. The date range was set as January 1 to December 31 of each year starting in 2015. Before that date, there were simply too few results (sometimes none). The earliest date range was thus set to January 1, 2000 to December 31, 2014. No omitted results were available for this period. For all search engines, the region was set to "Any Region" and the language to "English".

3.3 Extraction of URLs

All searches were conducted in Nancy (France) on July 19th 2026, with the Chrome web browser. URLs and their metadata (name of website, title of page, snippet of its content, etc.) were retrieved using the Instant Data Scraper browser extension[4], an automated data extraction tool operated by Flavr Technology that extracts structured data from web pages and exports it in a tabular from, as .csv or .xlsx files. The tool offers a crawling function that extracts several pages of results. However, there were some issues: the crawling was sometimes interrupted; and it was simply impossible to use with Yandex Search (the search results were simply copied and pasted in a .csv file).

3.4 Annotation of search results

The various .csv files generated by the Instant Data Scraper extension were combined in R (version 4.5.1). Overall, the searches generated 2309 search results; there is a lot of overlap, as was to be expected, with only 1604 unique URLs. Less expected, maybe, is the extent of the difference between the number of unique hits returned by DuckDuckGo (n=90) and Yandex (n=56) on one hand, and Google on the other (n=1512). The results were then manually annotated to asses their quality. The first step of the annotation process consisted in determining the date of each page. For Reddit and Quora pages, which account for a large portion of the results (see Section 3.7), this involved specific measures. Since dates of posts are not visible on Reddit's default interface, the old interface ("Old Reddit") was viewed by replacing "www" by "old" in URLs. For Quora, it was necessary to create an account and to add "/log" at the end of URLs to view a post (or "question")'s full history, including its publication date and answers. In some rare cases, such as dictionary entries and encyclopedia articles, no date was available.

Then, each page was inspected to determine if it was relevant, i.e., (1) if it contained the phrase "female as a noun" and (2) the phrase was used in a metalinguistic comment or discussion. This process was far from straightforward. For example, for Facebook results, it was often necessary to scroll far down a comment section to view the relevant comment(s), if they existed. In Reddit discussion threads, comments are sometimes hidden and must be "loaded" by clicking on a link, or can be viewed by accessing the pages in JSON format by appending ".json" to the end of URLs. Given the tediousness of the task, I decided to restrict the analysis presented in this paper to pages published until 2019 included (283 unique URLs). While this leaves out the major portion of the dataset, it still allowed me to meet the main objective of the study by focusing on earlier discussions.

All results considered not relevant didn't contain the search expression "female as a noun" – or, the search expression did not appear in the main discussion or article, but somewhere else on the page (for example, in "related posts" or "related topics" on discussion forums). This was frequent in Reddit and Quora results. The search expression always featured in metalinguistic discussions; thus, all pages that contained "female as a noun" were annotated as relevant. 167 results were annotated as relevant and 117 as not relevant.

3.5 Comparison between search engines

Combined, the four types of Google searches provided the bulk of relevant results (150/167 pages). Yandex only returned 20 relevant pages, and DuckDuckGo 16. Of the 16 pages not detected by Google, 8 were returned by Yandex and 8 by DuckDuckGo, with no overlap between the two search engines. 12 of these 16 pages not detected by Google were discussion threads posted on discussion forums, mostly Reddit but also smaller forums. Figure 1 represents the distribution of results by year; it only shows the results detected by each search engine and not by the others (i.e., the "Yandex" category shows the results returned by Yandex but not by Google and DuckDuckGo). It shows that very few results, or no result at all, were returned

[4] https://chromewebstore.google.com/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah?hl=en

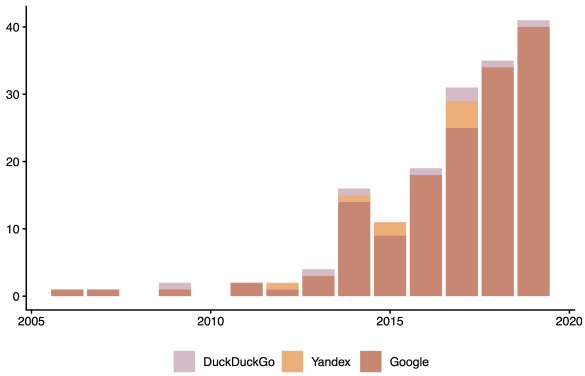


Figure 1: Distribution of relevant results detected exclusively by each search engine (pre-adjustments made in section 4.1)

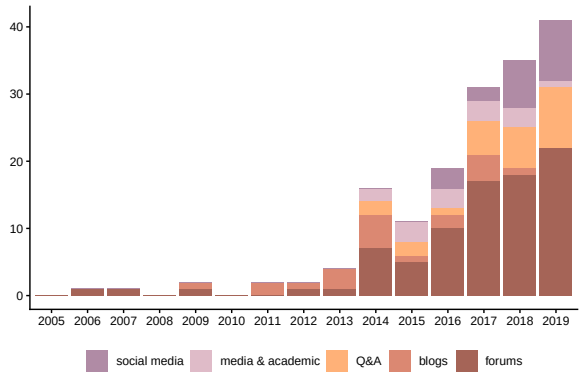


Figure 2: Composition of the dataset (post-adjustments)

for the years 2005-2013, and that most of these were returned by Google searches.

3.6 Comparison between Google searches

Google searches are compared for the years 2015 to 2019 (because a single combination of parameters was used for years prior to 2015). By far, the searches performed with a yearly time range and including omitted results returned the most results (n=93), followed by the global search with omitted results (n=43). Searches that didn't include omitted results yielded fewer results (n=35 with time range, n=39 without).

3.7 Composition of the dataset

The dataset is heterogeneous, with pages from 53 websites (social media, blogs, the Q&A website Quora, forum, media and academic websites). Forums, which are the earliest sources, constitute about half of the dataset (84 pages, or 50.3%). Most forum pages come from Reddit (55 pages, 32.93% of dataset). Quora (25 pages, 14.97%) and Facebook (16 pages, 9.58%) are the other two major sources of data. The distribution of data sources per year (Figure 2) reflects sociocultural shifts in internet use, with the decrease in blog content and the rise of social media. Despite Twitter popularity, only two results originate from the microblogging website, which may be due to recent changes affecting Twitter/X pages' indexing by Google (Peters, 2023).

4 Metalinguistic discourse analysis

4.1 Adjustments made to the dataset

The dataset analysed in this section differs slightly from the dataset presented above. Some results considered relevant had to be excluded because they were actually duplicates: for example, in forums, "female as a noun" sometimes appeared on several different pages (and thus URLs) of a single discussion thread. Here, several pages of a same discussion thread are considered a single observation. In one instance, a discussion thread spanned two different time intervals (the years 2006 and 2012); since the study focus on the diachronic developments of the "female" metalinguistic discourse, I considered the sets of comments posted during these two years as two different observations. In one case, an interview was published on two different websites; I only retained one occurrence. Only textual content was included in the dataset, meaning that two videos (posted on Facebook and Instagram) were excluded. Furthermore, the dataset was expanded by performing searches within websites using their search function when available; and by including web pages linked by internet users in forums and articles, some of which no longer existed and had to be retrieved using the Internet Archive Wayback Machine[5]. Overall, 12 new web pages were added to the dataset. The final dataset contains 167 observations, dated 2001 to 2019. It is available on OSF (link); it contains links to pages, websites names, dates, and annotations (but no verbatim comments).

4.2 The framing repertoire

The framing repertoire of "female" as a noun was constituted by manual annotation of metalinguistic comments by a single annotator (the author). Because the goal was to identify the set of arguments used rather than counting their frequency on each page, each discursive space was treated as a

[5] https://web.archive.org/

whole, with arguments found in a discussion thread or comment section documented only once, even when there were repeated by several people. Two types of framing were explored:

- The arguments mobilised by internet users to criticise and reject the usage of "female" as a noun.
- The social categories attached to the use of the noun; or, in other words, according to critics, who typically use "female" as a noun?

4.2.1 Arguments

Arguments were identified through a two-step process, beginning with the extraction of qualifiers, and then the coding of these qualifiers into distinct argumentative categories. 16 different arguments were identified; they are presented in Table 1, with examples (paraphrased, as well as all excerpts from the corpus cited in the paper, so that authors are not identifiable), the frequency of each category, and the year it first appears in the corpus. Some arguments are semantically quite similar (e.g. "dehumanising" and "for animals"; "derogatory" and "misogynistic"); they were retained as separate categories to capture the nuances of metalinguistic discourse. Because of the non-exhaustive and biased nature of the corpus, the exact date or origin of each argument in online discourse can obviously not determined. However, what Table1 shows is that, between 2015 and 2019, and despite the much larger number of pages in that period, no new arguments were introduced in the corpus. This suggests that, by 2015 (and at least until 2019), the metalinguistic debate had solidified and possibly reached a point of saturation.

4.2.2 Social categories

Descriptors referring to "typical" users of "female" as a noun, according to the authors in the corpus, were manually extracted in statements like " It's often used by incels which is disgusting" [Reddit, 2018] or "it makes you sound like Ferengi". [The Straight Dope Message Board, 2015]. Nuances were preserved, specifically in descriptors referring to misogynistic communities or movements, such as pick-up artists, MGTOW (Men going to their own way), MRAs (Men rights activists), red-pillers, etc. Table 2 shows the descriptors that appear at least twice in the corpus, with their overall frequency and year of first mention. Descriptors referring to misogynistic communities (incels, MRAs, etc.) or individuals (misogynists, anti-feminists), and figures of hegemonic masculinities (bros, neckbeards, nerds/geeks, "nice guys"; cf Ensmenger, 2015), highlighted in bold, constitute the majority of mentions (96/162, or 59.26%). "Ferengi" refers to the extraterrestrial species depicted in the American science fiction franchise *Star Trek*, who are characterised by their extremely capitalistic and sexist social norms and who systematically refer to women as "females". Not only is "Ferengi" the most frequent and earliest descriptor, but metalinguistic comments also frequently feature images of Ferengi or references to their pronunciation of the word: "They objectify us. We're just 'Feeeeeeemaaaales'. (Picture a Ferengi saying it)" [Reddit, 2013]. This suggests that references to Ferengi have become a staple of metalinguistic discourse of the noun "female, which is supported by the fact that the subreddit r/MenAndFemales (created in 2020 and devoted to calling out uses of "female" as a noun) features an image of a Ferengi as its profile image.

The use of "female" by the other social categories presented in Table 2 is not always seen as a derogatory. Metalinguistic commentary is often nuanced, recognising the importance of context: "In a scientific conversation [...] then sure, "female" as noun is fine. But used in a casual conversation, as it often is these days, it carries an underlying implication of otherness." [Reddit, 2017]. It also acknowledges other types of sociolinguistic variation linked to professional or linguistic background: "If the speaker isn't a native English speaker, I'm more inclined to give them the benefit of the doubt." [Quora, 2018]; "If he has a military or law enforcement background, I tend to give him the benefit of the doubt. ". [The Straight Dope Message Board, 2012].

Finally, like with arguments, the analysis shows that, despite the limited size of the corpus before 2014, the majority of descriptors are already present by that date: ideologies connected to the use of "female" as a noun are already crystallized in metalinguistic discourse.

5 Conclusions

This study used three search engines as points of entry to the web as corpus, in order to constitute and analyse a corpus of metalinguistic comments about the noun "female". The comparison between results provided by Google, Duck-

Argument	Examples	1st mention	Freq.
for animals	A kangaroo of the fairer sex is a female, but a human with two X chromosomes is a girl or a woman. [The Straight Dope Message Board, 2001]	2001	60
asymmetry	Men and Women is okay. Males and Females too. Men and Females... stupid [Resetera forum, 2017]	2001	43
weird	Referring to a woman using female as a noun is just weird [Ask Metafilter, 2007]	2001	37
dehumanising	As a rule of thumb, I never use "male" or "female" as a noun when describing people because "man" or "woman" is a lot less dehumanizing. [Reddit, 2019]	2006	79
derogatory	"female" is never followed with a positive comment [Resetera forum, 2017]	2006	52
sex vs. gender	For humans, sex and gender are two distinct concepts. A person can identify as a woman without being biologically female, and vice versa. [Reddit, 2014]	2006	16
misogynistic	I think using "female" as a noun to refer to women should be banned. [...]. Guys who do that are usually just misogynists. [The Snark Theater blog, 2018]	2009	33
a slur	It seems like OP is using the word "female" because she can't say "slut." [Ask A Manager comment section, 2014]	2009	9
reduces women to reproductive functions	I think that 'female' is outdated. Also, it reduces women to their reproductive capacity in contexts where it plays little to no role. [Media Diversity Institute website, 2019]	2011	25
a red flag	I have no idea why, but people using female as a noun sends up red flags in my mind. [Quora, 2019]	2011	18
distancing	I feel like they do this to distance themselves from us, to define women as the "other" [Reddit, 2017]	2012	40
grammatically incorrect	Female is an adjective, not a noun. Woman is a noun. [Reddit, 2018]	2012	32
transphobic	Not all women are biologically female, and equating "female" with "woman" invalidates gender-nonconforming and trans individuals. [Buzzfeed, 2014]	2013	15
a dog whistle	Now it's a bit of a dog-whistle. When a man uses it, it's often to disparage women [Reddit, 2014]	2014	1
a sign of discomfort with "woman"	Using "female" as a noun also suggests discomfort with the word "woman" as if it were indecent or inappropriate.. [Reddit, 2018]	2015	4
racist	In my experience people who use "females" in a derogatory way only reinforce the misogynoir that Black women face. [The Straight Dope Message Board, 2015]	2015	3

Table 1: Arguments used by critics of the noun "female"

DuckGo, and Yandex shows, unsurprisingly, that Google vastly outperforms other search engines in terms of number of relevant results returned. However, the few results provided by DuckDuckGo and Yandex (and not by Google) were valuable for the present study; specifically, they

Descriptors	1st mention	Freq.
Ferengi	**2006**	**27**
men terrible with women	**2007**	**3**
law enforcement	2011	13
misogynists	**2011**	**15**
the military	2011	24
bros	**2012**	**2**
medical/science professionals	2012	18
MRAs	**2012**	**9**
neckbeards	**2012**	**7**
nerds/geeks	**2012**	**4**
weird guys/douchebags	**2012**	**3**
Black men	2013	6
pick-up artists	**2014**	**5**
Hispanic men	2015	2
incels	**2017**	**13**
"nice guys"	**2018**	**3**
anti-feminists	**2018**	**3**
ESL speakers	2018	3
MGTOW	**2018**	**2**

Table 2: Social categories associated with the use of "female" as a noun

included results from smaller forums (like citydata.com, english.stackexchange.com, italki.com and neogaf.com), which helped diversify the corpus. Similarly, the comparison between various Google searches showed that it can be useful to use various combinations of parameters in order to get as many relevant results as possible. Choosing the option "repeat the search with the omitted results included" and setting a time range improved the performance. One-year intervals were used as the time range and it is possible that using shorter time spans would have yielded even more results.

While the corpus created provides a partial and biased view because of the limitations of search engines and of the methodology used, its analysis provides new insights on the "female" debate. First, it shows that the new "female" debate builds on arguments put forward by earlier critics of the noun: its dehumanising and derogatory connotations, and the animal comparison. When allowances are made, for example in medical and scientific discourse, it is only under the condition that usage is symmetric ("males and females"), a requirement also specified by style guides (for ex., Miller and Swift, 1980; Froke et al., 2022; American Psychological Association, 2010). Other arguments ("female" is racist, transphobic, and does not refer to gender) reveal the influence of feminist and transgender activism. The corpus also contains newer arguments, which go beyond noting pejorative connotations: some internet users see "female" as a slur, not synonymous with "woman" but with "bitch" or "slut". The "red flag" and the "dog whistle" arguments implicitly tie the use of the noun to misogynistic individuals and communities. The analysis of "typical" users of "female" shows that the pejoration of the term (and possible increase in usage) is connected to the masculinist ideologies of the manosphere. The linguistic argument (that "female" is an adjective and not a noun, and thus using it as a noun is grammatically incorrect) is also new. It is sometimes countered with references to dictionary entries. Some users invert the critique, arguing that widespread misuse of "female" has led dictionaries to include the noun: "The misuse of this word has become so widespread that some dictionaries now list "noun" as a secondary meaning. That doesn't make it acceptable. Stop. This. Right now. . "[Facebook, 2019]. The prescriptive force of this new norm exemplifies what Deborah Deborah Cameron (2012) terms "verbal hygiene", a form of metalinguistic discourse that, while nominally about grammar, reflects "deeper anxieties which are not linguistic, but social, moral and political" (p. vii). Finally, looking diachronically at the corpus reveals that, despite the small number of pages in the earlier years (2001-2015), by 2015, the framing repertoire of the debate has mostly been constituted, with no new arguments emerging. It is possible, however, that new arguments and ideological stances have emerged since 2019. Exploring this will be the focus of my future research.

Limitations

The nature of search engines makes this research unreproducible and the corpus fundamentally biased: it is impossible to know how many pages were filtered by search engines, and why. Other limitations are the facts that using search engines do not give access to websites that have disappeared and pages or comments that have been deleted (unlike archives), and that the dataset was annotated by a single annotator. Moreover, the search expression "female as a noun" was deliberately chosen for its limited scope, with the intent to minimise false positives. This means that metalinguistic discus-

sions which do not include the expression "female as a noun" were not included in the corpus; it also means that the relatively limited number of search engine page results made it possible to manually examine all pages published in a 19-year period, which may not be realistic for studies focusing on more frequent words or expressions.

Acknowledgments

This research was funded by the French National Research Agency as part of the project "ANR-24-AERC-0011-01".

References

American Psychological Association. 1995. *Publication manual of the American Psychological Association*, 4th ed edition. Washington, DC.

American Psychological Association. 2010. *Publication manual of the American Psychological Association*, 6th ed edition. Washington, DC.

Paul Baker. 2014. *Using corpora to analyze gender*. Bloomsbury, London.

Sarah Banet-Weiser and Kate M Miltner. 2016. Masculinitysofragile: Culture, structure, and networked misogyny. *Feminist media studies*, 16(1):171–174.

Dennis E Baron. 1986. *Grammar and gender*. Yale University Press.

BBC News. 2009. Microsoft and Yahoo seal web deal.

Deborah Cameron. 2012. *Verbal hygiene*. Routledge, London; New York.

Deborah Cameron. 2016. To gender or not to gender? (thoughts prompted by the death of Zaha Hadid).

Margaret Doyle. 1995. *The A-Z of non-sexist language*. The Women's Press.

Nathan Ensmenger. 2015. "Beards, sandals, and other signs of rugged individualism": masculine culture within the computing professions. *Osiris*, 30(1):38–65.

Bergen Evans and Cor Evans. 1957. *A dictionary of contemporary American usage*. Random House, Inc.

Andrew Ferley. 2022. *The semantic derogation of female*. Master's thesis, York University.

Marie Flesch and Sophie Bailly. 2020. *Mapping the itineraries and interests of cisgender and transgender internet users with the RedditGender corpus*, page 93–112. L'Harmattan.

Henry Watson Fowler. 1926. *A dictionary of modern English usage*. Clarendon Press, Oxford.

Paula Froke, Anna Jo Bratton, Andale Gross, Jeff McMillan, Pia Sarkar, Jerry Schwartz, and Raghuram Vadarevu. 2022. *The Associated Press stylebook 2022-2024*, 56 edition. The Asssociated Press, New York.

Marilyn Frye. 2005. Categories in distress. In Barbara S. Andrew, Jean Keller, and Lisa H. Schwartzman, editors, *Feminist interventions in ethics and politics: feminist ethics and social theory*, page 41–58. Rowman Littlefield Publishers.

Azadeh Hajian Hoseinabadi and Mehrdad CheshmehSohrabi. 2024. Proposing a new combined indicator for measuring search engine performance and evaluating Google, Yahoo, DuckDuckGo, and Bing search engines based on combined indicator. *Journal of Librarianship and Information Science*, 56(1):178–197.

John Lesslie Hall. 1917. *English usage: Studies in the history and uses of English words and phrases*. Scott, Foresman.

Marlis Hellinger. 2010. Guidelines for non-discriminatory language use. In Ruth Wodak, Paul E Kerswill, and Barbara Johnstone, editors, *The SAGE handbook of sociolinguistics*, page 565–582. SAGE Publications. Publisher: Sage.

JenJens. 2026. Is the incel website content too much?

Santeri Joensuu. 2023. *The plural "females" and its derogatory use. A corpus study on the sexist aspects in the use of the word*. Master's thesis, Tampere University.

Andrew Kehoe. 2020. Web corpora. In Magali Paquot and Stefan Th. Gries, editors, *A practical handbook of corpus linguistics*, page 329–352. Springer.

Robin Tolmach Lakoff. 1975. *Language and woman's place*. Harper Row.

LC. 2018. "The best human benefactor of WOMAN", Matthew Vassar and Sarah Josepha Hale.

Dirk Lewandowski. 2023. *Understanding search engines*. Springer.

Shahan Ali Memon and Jevin D West. 2024. Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable. *arXiv preprint arXiv:2402.11707*.

Casey Miller and Kate Swift. 1980. *The handbook of nonsexist writing*, 1st edition. Lippingcott Crowell, New York.

Sara Mills and Louise Mullany. 2011. *Language, gender and feminism: Theory, methodology and practice*. Routledge.

Amanda Montell. 2024. *Wordslut: a feminist guide to taking back the English language*. Taylor Francis.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. 2018. Building a web-scale dependency-parsed corpus from commoncrawl. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Jay Peters. 2023. Tweets aren't showing up in Google results as often because of changes at Twitter.

Jessica R Peters, Allison Stumper, Katja M Schmalenberger, Andy J Taubman, and Tory A Eisenlohr-Moul. 2023. Improving rigor through gender inclusivity in reproductive psychiatric science. *Comprehensive Psychoneuroendocrinology*, 16:100194.

Amrit Poudel, Yifan Ding, Tim Weninger, and Jürgen Pfeffer. 2025. Digital gatekeepers: Google's role in curating hashtags and subreddits. In *Proceedings of the 63rd annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, page 2402–2415.

Melissa Aubrey Robinson. 2018. *A man needs a female like a fish needs a lobotomy: The role of adjectival nominalization in pejorative meaning*. Master of arts thesis, University of North Texas.

William Safire. 2007. Language: Woman vs. female. *The New York Times*.

Ariel Salleh. 1992. The ecofeminism/deep ecology debate. *Environmental Ethics*, 14(3):195–216.

Muriel R. Schulz. 1990. The semantic derogation of women. In Deborah Cameron, editor, *The feminist critique of language: A reader*, 1st edition, page 135–147. Routledge.

Robert Sigley and Janet Holmes. 2002. Looking at girls in corpora of english. *Journal of English Linguistics*, 30(2):138–157.

John Sinclair. 2005. *Corpus and text - Basic principles*. Oxbow Books, Oxford.

statcounter. 2026. Search engine market share worldwide.

The University of Chicago Press Editorial Staff. 2024. *The Chicago Manual of Style*, 18 edition.

Marko Ćurković and Andro Košec. 2018. Bubble effect: including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility. *BMC medical research methodology*, 18(1):130.

Keeping the Ums: Conversation-Analysis-Informed Verbatim ASR for Web Speech

Steven Coats
English, Faculty of Humanities, University of Oulu
90014 University of Oulu, Finland
steven.coats@oulu.fi

Abstract

Web video makes massive spoken-language corpora possible, but automatic speech recognition (ASR) routinely normalizes away fillers, repetitions, repairs, and hesitations. Because Conversation Analysis (CA) transcripts meticulously document these interactional phenomena, they offer an ideal yet underutilized resource for ASR adaptation. We show that fine-tuning Whisper large-v3 on just 4.47 hours of CA-annotated Australian and US conversation substantially improves verbatim transcription on heterogeneous web speech without sacrificing lexical accuracy. On a 57-clip benchmark from 39 Australian and New Zealand local-government channels, fine-tuning reduces verbatim word error rate from 16.06% to 11.39% (a 29% relative reduction) and raises filler recovery (F_1) from .117 to .819. These gains transfer directly to New Zealand speech absent from training (reducing verbatim error from 18.24% to 11.71%), demonstrating that small CA datasets teach the model a generalizable transcription policy rather than narrow acoustic memorization. We argue that targeted verbatim adaptation should become standard practice in web-as-corpus pipelines to ensure large-scale resources like CoANZSE preserve the interactional evidence needed for linguistic research.

1 Introduction

The web has made it possible to build speech corpora at a scale that was once impractical. CoANZSE Audio, for example,[1] provides searchable access to more than 195 million words from videos uploaded to YouTube by Australian and New Zealand councils and other local-government entities (Coats, 2024). Such resources can support work on regional variation, morphosyntax, phonetics, discourse, and interaction. Their linguistic value, however, depends on what the transcript represents.

[1] https://coanzse.org

Modern ASR systems are often optimized for readable text, a design philosophy which has been characterized as emphasizing "clarity of intent" (Lea et al., 2023). Systems such as Whisper (Radford et al., 2023) often omit *uh*, *um*, repeated words, and false starts, and normalize non-standard inflectional forms, making transcripts less useful for researchers studying dialectal speech or interactional features. While disfluencies, repetitions, or filler particles carry relatively little semantic content alone, they can play a crucial role in discourse. From an interactional perspective, filled pauses can signal production delay and upcoming complexity (Clark and Fox Tree, 2002), prompting listeners to anticipate new information (Arnold et al., 2003), or can be associated with dispreferred responses (Kendrick and Torreira, 2015) such as disagreement or rejection. From a sociolinguistic perspective, these items can exhibit systematic regional variation (Wieling et al., 2016) and evolve over time in the same manner as content words (Fruehwald, 2016; Tottie, 2011). For speech corpora, transcriptions which preserve filler particles, repetitions, and repairs (i.e. verbatim transcriptions) can therefore provide important evidence for the analysis of discourse content, interaction, and language variation and change. At web-corpus scale, transcription normalization can hinder dialectological, sociolinguistic, and interactional research, necessitating the development of models specifically designed to recover these features.

Recent work has addressed verbatim ASR directly. As might be expected, fine-tuning on in-domain transcripts in which fillers are carefully preserved improves recovery of these items (Horii et al., 2022; Lin et al., 2025), but many gold standard transcripts used as training data do not carefully preserve such features, and manual annotation is time-consuming and expensive. The importance of domain adaptation as part of the corpus-design pipeline is also increasingly recognized in corpus

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 192–199
October 29, 2026 ©2026 Association for Computational Linguistics

linguistics and research on varieties of English. Fine-tuning an ASR model with conversational Singapore English speech, for example, makes locally salient forms such as *lah*, *lor*, *sia*, and *kena* available for corpus analysis, whereas they are largely absent from the output of default ASR (Coats et al., forthcoming). Re-engineering parts of the ASR system such as the tokenizer and the procedures used to derive timing information can improve verbatim performance (Zusag et al., 2024). More extensive interventions allow controllable decoding: by prepending a prompt token (e.g., `<verbatim>` or `<intended>`) to the decoder, a single model can be instructed either to transcribe literal hesitations and repairs or to filter them out for readable text (Wagner et al., 2026). From the practical perspective of corpus linguists, however, finding an appropriate balance between training data selection, manual annotation, and system redesign is crucial. The practical question for a web corpus of national varieties of English remains open: can a small amount of existing, carefully transcribed conversation improve verbatim ASR on heterogeneous Australian and New Zealand web video without sacrificing lexical accuracy?

We address this question by adapting conversation-analytic and related interactionally detailed data available for research use. Conversation Analysis (CA) is an approach to recordings of talk-in-interaction in which detailed transcripts are used to analyze how participants organize social action through turn-taking, sequence organization, repair, overlap, and timing, typically through qualitative sequential analysis. These transcripts preserve precisely the hesitations, repetitions, cut-offs, and other features that general-purpose ASR frequently normalizes away, making them a potentially valuable but underused source of supervision for corpus construction.

We contribute (i) a context-assisted manual benchmark from 100 clips in a candidate source-recording pool for a future CoANZSE release, with a prespecified 57-clip quality-controlled primary subset; (ii) an evaluation of two Whisper large-v3 models adapted on this CA data, namely a verbatim-only full fine-tuned model and a controllable dual-policy model using Low-Rank Adaptation (LoRA; Hu et al., 2022), benchmarked against base Whisper and CrisperWhisper2 for lexical accuracy and the preservation of interactionally relevant phenomena; and (iii) evidence that only 4.47 hours of detailed conversational supervision transfers across recording domains and to New Zealand speech absent from training. Supplementary source-corpus experiments examine adaptation of CrisperWhisper2 and the effect of adding a third Australian corpus. Code, evaluation manifests, sanitized predictions, and scoring scripts are available in an anonymous repository.[2]

2 Data and method

2.1 Interactionally detailed training data

The starting point of our experiments comprised two corpora with interactionally detailed transcripts and the corresponding audio: the La Trobe Corpus of Spoken Australian English (LTCSAusE; Mullan, 2002) and the Santa Barbara Corpus of Spoken American English (SBCSAE; Du Bois et al., 2000–2005). LTCSAusE comprises conversational interaction recorded in Melbourne; SBCSAE includes diverse US conversational interactions. Both corpora preserve substantial conversational detail, including precise timing information and annotations for overlapping speech, pauses, pitch and prosody, hesitations and disfluencies, repetitions, and filler particles.

From the source transcripts, we normalized corpus-specific notation into a common lexical convention for fillers (e.g. *uh/er*, *um/erm*, *mm*), fragments (e.g. *d-didn't*), repetitions, and non-speech events (e.g. music and noise). Training examples were cut to exact acoustic boundaries and filtered to avoid substantial overlap and mixed-speaker target segments. The resulting split contains 798 training clips (4.47 h), 167 validation clips (1.10 h), and 210 held-out clips (1.24 h). Only the training and validation splits were used during model development.

For an additional data-ablation experiment, we prepared the Griffith Corpus of Spoken Australian English (GCSAusE; Haugh and Chang, 2013), a registered-access TalkBank corpus with CA-style transcripts and 40 recordings. The same normalization and boundary criteria yielded 192 non-overlapping, single-speaker clips (0.96 h): 124 training, 30 validation, and 38 test clips. Adding these items produced a three-source dataset with 922 training, 197 validation, and 248 test clips. Recording-level splits prevented material from the same source recording from crossing partitions.

Although limited in size by the amount of suitable data available, this dataset provides consistent

[2] https://github.com/stcoats/CAWhisperWaC13

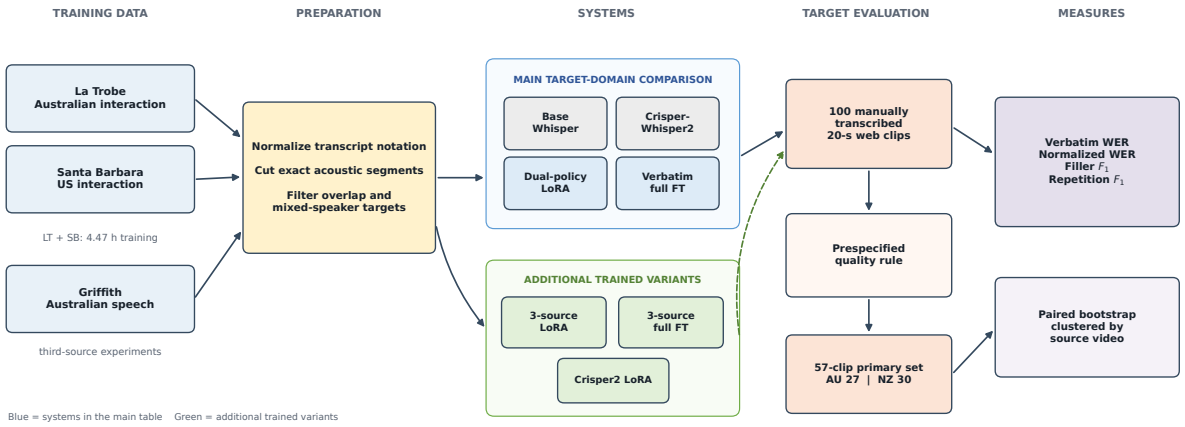


Figure 1: Experimental workflow for verbatim ASR adaptation and evaluation on candidate CoANZSE web speech.

positive supervision for features commonly omitted by general-purpose ASR and weak web captions. While CA transcripts also document fine-grained prosody, overlap, and temporal latching, the present study focuses specifically on recovering lexicalized interactional phenomena (fillers, repetitions, and repairs) for searchable text corpora, leaving acoustic timing and overlap modeling to future work. We use it to test whether such supervision transfers to heterogeneous Australian and New Zealand web speech.

2.2 Models

Verbatim full FT. We updated the complete encoder–decoder model of Whisper large-v3 against the verbatim target. Training used bfloat16 on four NVIDIA GH200 GPUs, an effective batch size of 16, a learning rate of 10^{-5}, 10% warmup, at most ten epochs, and early stopping with patience three. The selected checkpoint minimized verbatim validation WER.

Dual-policy LoRA. Adapting one of the techniques described in (Wagner et al., 2026), we used the CA-transcript-derived text as the verbatim target and created a corresponding intended-output version by deterministically removing fillers, repetitions, and other disfluencies (for example, transforming verbatim *"we uh we had to d-decide"* into intended *"we had to decide"*). The model was trained on both versions, with each audio–transcript pair marked by a new decoder-prefix token: either `<verbatim>` or `<intended>`. Training otherwise followed a standard PEFT/LoRA procedure. Adapters were applied throughout the encoder and decoder attention layers (rank 32, alpha 64, dropout .05). Because the original Whisper parameters remained frozen, the embeddings for the two new policy tokens were trained separately alongside the LoRA adapters.

This procedure differed from the staged method of Wagner et al. (2026), who first trained only newly introduced token embeddings and then unfroze the decoder while retaining a frozen encoder. Because our adapters operated in both the encoder and decoder, they could adapt both output policy and acoustic/domain representations; the experiment did not isolate these two effects. We evaluated only the verbatim mode in the main target-domain comparison.

Supplementary adaptation experiments. In addition to evaluating out-of-the-box CrisperWhisper2 as a baseline in our primary benchmark, we applied the same LoRA configuration directly to the CrisperWhisper2 large model in a supplementary experiment, using the La Trobe and Santa Barbara verbatim targets and CrisperWhisper2's verbatim decoder prefix. As in our dual-policy model, rank-32 adapters were applied to the query and value projections throughout both the encoder and decoder; the original CrisperWhisper2 parameters remained frozen. This experiment tested whether the specialist model would benefit from the same conversational supervision. Finally, we repeated verbatim full fine-tuning of Whisper large-v3 after adding GCSAusE. This three-source model used the same full-training hyperparameters and selected a checkpoint by the unweighted validation-WER macro across the three source corpora. It stopped after five epochs and restored the epoch-two checkpoint.

In the primary target-domain evaluation, we

compared both principal adapted systems (verbatim full FT and dual-policy LoRA) against unmodified Whisper large-v3 and out-of-the-box CrisperWhisper2. Each system received the identical mono 16-kHz target waveform and no textual context; decoding used standard greedy search (temperature 0).

2.3 Independent web-speech evaluation

The evaluation audio comes from original source recordings re-collected for a candidate future CoANZSE version; it is not yet part of the released CoANZSE v1.0 index. While published by local-government channels, the source recordings in CoANZSE span diverse communicative genres (Coats, 2024): beyond council meeting debates, they include public consultations, citizen interviews, community event presentations, and informational videos, capturing both formal deliberations and spontaneous conversational interaction. We selected 50 source videos from 50 local-government channels, 25 from Australia and 25 from New Zealand. From each video we took one 20-second clip whose YouTube caption contained a filler marker and one comparison clip without such a marker. YouTube captions were used only to stratify candidates, not as evaluation references.

Because conversational fillers and repair particles are frequently missed in high-volume crowdsourced annotation, we prioritized expert judgment over large manual transcript volumes. A single linguist (the author, a native English speaker trained in corpus linguistics) transcribed all 100 clips with access to extended acoustic context and video metadata. To resolve local names and low-context fragments, annotation was context assisted: the annotator could listen to a longer window and inspect the video title, council and location, and neighboring YouTube captions. The target transcript nevertheless describes only the exact 20-second clip. Annotation preceded evaluation and did not display system predictions.

Before inference, we froze a primary quality rule: intact boundaries, good or usable audio, and no more than minor overlap. Fifty-seven clips from 39 videos met the rule (3,188 scored words; 19 minutes): 27 Australian and 30 New Zealand, with 27 filler-marked and 30 unmarked candidates. We report all 100 clips as a sensitivity analysis.

2.4 Metrics and uncertainty

Verbatim WER (vWER) retains audible fillers, repetitions, fragments, contractions, and reduced forms. Normalized WER removes fillers and fragments, collapses adjacent repetitions, and normalizes selected reduced/full forms. Both ignore punctuation and case and normalize for inaudible orthographic choices, including Australian/UK versus US spelling, titles, initialism formatting, and macrons (common in some Māori borrowings). Reference `[unclear]` tokens mask one uncertain unit rather than forcing a possibly incorrect gold word.

We additionally report F_1 for fillers and adjacent repetitions. Confidence intervals use 20,000 paired bootstrap replicates clustered by source video, preserving dependence between the two sampled clips. The clean references contain 129 fillers, 25 adjacent repetitions, and six word fragments. Fragment results are consequently too sparse for a substantive comparison and are not included in the main table. Reporting both WER variants is essential: vWER measures fidelity to conversational form, whereas normalized WER asks whether a system still recognizes lexical content after policy-related differences are removed.

3 Results

3.1 Target-domain results

Table 1 reports the results on the quality-controlled CoANZSE candidate subset. All three verbatim-oriented systems produced lower vWER point estimates than base Whisper. The full verbatim model performed best numerically, reducing vWER from 16.06% to 11.39%: an improvement of 4.67 percentage points, or 29.1% relative. The paired difference from base had a 95% interval of -6.32 to -2.99 points. The full model was also the only system for which the normalized-WER improvement over base had a paired interval wholly below zero (-1.92 points, CI -3.50 to -0.39).

The filler and repetition measures reveal a somewhat different pattern. CrisperWhisper2 achieved the highest filler F_1 (.859), narrowly ahead of the dual-policy LoRA (.843) and full model (.819), while the dual-policy LoRA achieved the highest repetition F_1 (.800). Reporting filler precision and recall alongside F_1 shows that CrisperWhisper2's slight advantage is driven primarily by precision (.843 vs .782 for full FT), whereas recall remains comparable across all three adapted models (.853–

System	vWER ↓	95% CI	norm. WER ↓	filler P / R	filler F_1 ↑	repetition F_1 ↑
Whisper Large-v3	16.06	13.41–18.80	12.11	1.0 / .062	.117	.375
CrisperWhisper2	13.11	9.82–16.88	12.44	**.843 / .876**	**.859**	.724
Dual-policy LoRA	12.17	9.78–14.82	11.07	.833 / .853	.843	**.800**
Verbatim full FT	**11.39**	9.23–13.78	**10.19**	.782 / .860	.819	.710

Table 1: Results on the prespecified quality-controlled CoANZSE candidate subset (57 clips). WER values are percentages. vWER intervals are 95% percentile intervals from 20,000 paired bootstrap replicates clustered by source video.

System	Australia	New Zealand
Whisper Large-v3	13.71	18.24
CrisperWhisper2	11.23	14.86
Dual-policy LoRA	11.88	12.44
Verbatim full FT	**11.03**	**11.71**

Table 2: Descriptive vWER (%) by country on the clean subset (Australia: 27 clips; New Zealand: 30).

.876). Base Whisper had filler precision of 1.0 but recall of only .062: its eight predicted fillers were correct, but it recovered only 8 of the 129 fillers in the reference transcripts. The three verbatim-oriented systems recovered substantially more of these conversational events.

The paired contrasts nevertheless do not identify a universal winner. The full model's vWER was 0.78 points lower than that of the dual-policy LoRA (95% CI -1.77 to 0.07) and 1.73 points lower than that of CrisperWhisper2 (CI -4.11 to 0.25), but neither interval excluded zero. Its normalized WER was lower than that of the dual-policy LoRA by 0.88 points (CI 0.03–1.88) and lower than that of CrisperWhisper2 by 2.24 points (CI 0.03–4.96). Thus, strong recovery of conversational events did not necessarily coincide with the most accurate lexical transcript.

The pattern was not confined to Australian speech. On the 30 New Zealand clips, vWER fell from 18.24% for base Whisper to 11.71% for the full model and 12.44% for the dual-policy LoRA, although neither adaptation corpus contained New Zealand speech. These subgroup results are descriptive because of the small sample.

A typical sequence of omissions by base Whisper is shown below (irrelevant words elided). The full model recovered the repeated articles and all three filler positions, although it rendered the first *uh* as *um*.

> **Reference:** "the **the** council's not … seek to **uh** replace … agencies **um** in that regard **uh** … the **the** proponent …"
>
> **Base Whisper:** "the council is not … seek to replace … agencies in that regard … the proponent …"
>
> **Full FT:** "the **the** council's not … seek to **um** replace … agencies **um** in that regard **uh** … the **the** proponent …"

A sensitivity analysis using all 100 annotated clips preserved the vWER ranking: 19.55% for base Whisper, 16.25% for CrisperWhisper2, 14.07% for the dual-policy LoRA, and 13.84% for the full model.

The same ranking appeared in both sampling strata. For base Whisper, CrisperWhisper2, the dual-policy LoRA, and the full model respectively, vWER was 15.05%, 12.17%, 11.19%, and 10.93% on the 27 filler-marked clips, compared with 16.99%, 13.98%, 13.07%, and 11.81% on the 30 nominally unmarked clips. Audible fillers occurred in both strata: YouTube captions were useful for retrieving likely examples, but the absence of a captioned filler was not a reliable negative label.

3.2 Supplementary source-corpus experiments

Table 3 compares the two- and three-source full models on all three held-out source-corpus partitions. Adding 0.64 hours of GCSAusE training audio produced small and mixed changes. It lowered vWER on La Trobe from 14.84% to 14.41% and on GCSAusE from 30.29% to 28.87%, while SBC-SAE vWER increased from 9.97% to 10.20%. The unweighted three-corpus macro fell from 18.37% to 17.83%. GCSAusE repetition F_1 increased from .400 to .545, but filler F_1 changed only from .408 to .426 and declined slightly on the other two corpora. Additional data therefore produced modest redistribution rather than a uniform improvement.

Adapting CrisperWhisper2 with LoRA similarly improved lexical accuracy without improving every conversational-event measure. On held-out La Trobe, vWER fell from 20.29% to 17.07%, while filler F_1 fell from .757 to .599. On Santa Barbara, vWER fell from 14.11% to 10.65%, while filler F_1

Held-out corpus	System	vWER ↓	filler F_1 ↑	repetition F_1 ↑
La Trobe	Two-source full FT	14.84	.800	.709
	Three-source full FT	**14.41**	.793	**.719**
Santa Barbara	Two-source full FT	**9.97**	**.618**	.695
	Three-source full FT	10.20	.594	**.703**
Griffith	Two-source full FT	30.29	.408	.400
	Three-source full FT	**28.87**	**.426**	**.545**

Table 3: Held-out source-corpus results for the data-addition experiment. The two-source model used La Trobe and Santa Barbara; the three-source model additionally used GCSAusE. Values are descriptive point estimates.

fell from .494 to .430. Thus, applying the same adaptation data and method to a specialist verbatim model did not simply amplify its original strengths.

Finally, as a policy-control check, we decoded the held-out La Trobe and Santa Barbara speech using both modes of the dual-policy model. Relative to verbatim mode, intended-mode decoding reduced normalized WER from 19.60% to 13.62% on La Trobe and from 11.65% to 9.26% on Santa Barbara. This confirmed that the prefix tokens affected the model's output policy. The intended references were derived from the verbatim targets rather than independently annotated; this comparison therefore served as a manipulation check, not as an evaluation of intended-transcription quality. Furthermore, decoding the primary 57-clip CoANZSE test set in `<intended>` mode effectively suppressed reference fillers ($F_1 = .045$, predicting only three reference fillers) while achieving a normalized WER of 10.94%, statistically indistinguishable from verbatim mode (11.10%; 95% CI $[-1.13, +0.70]$), confirming policy controllability on the target domain without a lexical accuracy penalty on remaining words.

4 Discussion

The findings support a data-centric approach to web speech. Web captions are useful for discovery and rough lexical access, but they are not neutral representations of interaction. In our nominally "unmarked" comparison stratum, manual listening found genuine fillers that YouTube captions had omitted; the verbatim systems correctly recovered many of them. Caption absence should therefore not be treated as a negative label.

This is not a conventional same-domain adaptation setting. Supervision comes from research recordings of naturally occurring Australian and US interaction, whereas evaluation uses heterogeneous institutional web video; more than half of the clean test clips are New Zealand speech. The particularly large New Zealand reduction suggests that detailed supervision can transfer a transcription policy without requiring a matched corpus for every target variety. Accent adaptation and policy adaptation remain entangled in our encoder–decoder training, however, so the country split should be interpreted as transfer evidence rather than an accent-specific causal result.

The strongest production choice depends on purpose. CrisperWhisper2 is best if filler recovery alone dominates. The dual LoRA provides explicit verbatim/ intended control and the best repetition F_1 in this sample. The full model is the best candidate for rebuilding a searchable corpus because it combines high event recall with the lowest verbatim and normalized WER. In a corpus such as CoANZSE, that balance matters: hallucinated or substituted content damages lexical searches, while deleted disfluencies make interactional searches impossible.

For corpus construction, these errors have different downstream costs. A content substitution creates false positives and false negatives in ordinary lexical searches; deletion of a genuine filler or repetition removes the event from the corpus altogether and prevents it from entering an interactional sample. A production pipeline should therefore retain both a verbatim layer and a normalized search view rather than forcing one transcript to serve both purposes. The present scoring framework operationalizes that distinction even when the deployed model emits only verbatim text.

No single metric captures these requirements. Normalized WER makes base Whisper appear closer to the specialist systems by removing many systematic omissions before scoring; filler F_1 favors CrisperWhisper2 while concealing its greater lexical error. The paired results favor the full model as the current production candidate, but its modest vWER advantage over the dual LoRA is not

conclusive. CrisperWhisper2 achieved the highest filler F1 in our evaluation and its remit addresses a considerably broader problem than the present study, including controllable transcription policy, paralinguistic events, and word-level timing in multilingual data. Our experiment instead considered a narrower practical scenario: whether existing detailed conversational transcripts can improve the transcription of a particular regional web corpus using a comparatively conventional adaptation pipeline. The lower lexical accuracy of CrisperWhisper2 on CoANZSE may reflect differences between its training data and the Australian and New Zealand target domain (CrisperWhisper2 was trained in part on CORAAL, comprising African-American speech), but the present evaluation cannot isolate the contribution of accent, recording conditions, model configuration, or individual training corpora. The systems should therefore be understood as emphasizing different, potentially complementary forms of adaptation.

The supplementary experiments reinforce this interpretation. LoRA adaptation of CrisperWhisper2 reduced source-corpus WER but also reduced filler F_1, while adding the small Griffith dataset to full fine-tuning improved some corpus–metric combinations and slightly worsened others. Fine-tuning did not operate as a simple accumulation of improvements: the composition and transcription conventions of the supervision affected which properties were retained. This is particularly relevant when the intended output is a multi-purpose linguistic corpus rather than a single benchmark score.

With regards to an update of the CoANZSE resource, a potential v2 could preserve verbatim ASR as its archival annotation while deriving a normalized token layer for ordinary concordance queries. Users could then search for lexical content or interactional forms and inspect the same aligned audio. Unlike a single "clean" transcript, this design would not irreversibly remove conversational events; the transcription policy could be made explicit in the corpus documentation.

5 Conclusion

Base Whisper's readable output systematically removes conversational evidence needed for some kinds of linguistic research. A few hours of carefully segmented conversation-analysis data were sufficient to reduce verbatim error on independent Australian and New Zealand web speech while retaining strong lexical accuracy. For web speech corpora, transcription policy should be an explicit design and evaluation decision, not an undocumented side effect of the ASR model.

Limitations

Our evaluation focuses on a targeted, context-assisted benchmark of 100 annotated clips across 50 municipal channels (analyzed as a frozen 57-clip primary subset and a 100-clip sensitivity set). Reference transcripts were produced by a single annotator without inter-annotator agreement measurement, and results are situated within Australian, New Zealand, and US varieties of English. While event sparsity for rare categories like word fragments ($N = 6$) precludes firm claims for those specific subtypes, paired cluster-bootstrap resampling across 20,000 replicates demonstrates that the main reductions in verbatim WER and filler recovery are statistically robust across diverse source channels. Original web recordings can be stereo, whereas all systems were compared on a common mono 16-kHz conversion. The intended targets used by the dual model are deterministic transformations rather than independent human transcripts. Our encoder–decoder LoRA does not isolate output-policy learning from acoustic adaptation; a decoder-only or token-embedding-only ablation would do so more directly. We did not evaluate diarization or word timing, so our conclusions concern transcription only.

Ethical considerations

The benchmark is derived from publicly available institutional web video, but public availability does not eliminate privacy, copyright, platform, or contextual-integrity concerns. The audio was not originally recorded under a research-consent protocol. We therefore plan to release code, scoring tests, source identifiers, time offsets, and aggregate results, while redistributing audio, reference text, or trained weights only where source-corpus licenses and web-platform conditions permit. The benchmark material comes from a candidate future CoANZSE pool and should not be represented as part of the current public CoANZSE v1.0 index.

Acknowledgments

This work was supported by the European Union NextGenerationEU instrument and was funded by

the Research Council of Finland under grant number 358720. Computational resources were provided by Finland's Centre for Scientific Computing (CSC).

References

Jennifer E. Arnold, Maria Fagnano, and Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32(1):25–36.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Steven Coats. 2024. Building a searchable online corpus of Australian and New Zealand aligned speech. *Australian Journal of Linguistics*, 44(2–3):261–277.

Steven Coats, Carmelo Alessandro Basile, Cameron Morin, and Robert Fuchs. forthcoming. Fine-tuning ASR for corpus linguistics: Singapore English. *Digital Scholarship in the Humanities*.

John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. Santa Barbara Corpus of Spoken American English, Parts 1–4.

Josef Fruehwald. 2016. Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, 22(2):41–49.

Michael Haugh and Wei-Lin Melody Chang. 2013. Collaborative creation of spoken language corpora. In Tim Greer, Donna Tatsuki, and Carsten Roever, editors, *Pragmatics and Language Learning*, volume 13, pages 133–159. National Foreign Language Resource Center, University of Hawaii at Manoa, Honolulu, HI.

Koharu Horii, Meiko Fukuda, Kengo Ohta, Ryota Nishimura, Atsunori Ogawa, and Norihide Kitaoka. 2022. End-to-End Spontaneous Speech Recognition Using Disfluency Labeling. In *Interspeech 2022*, pages 4108–4112.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kobin H. Kendrick and Francisco Torreira. 2015. The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4):255–289.

Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Jhen-Ke Lin, Hao-Chien Lu, Chung-Chun Wang, Hong-Yun Lin, and Berlin Chen. 2025. Acoustically Precise Hesitation Tagging Is Essential for End-to-End Verbatim Transcription Systems. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 163–166.

Kerry Mullan. 2002. The La Trobe corpus of spoken Australian English.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Gunnel Tottie. 2011. Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, 16(2):173–197.

Laurin Wagner, Mario Zusag, and Bernhard Thallinger. 2026. Transcription policy as a latent variable: Activating controllable verbatim ASR with word-level timing. *Preprint*, arXiv:2607.18934. Accepted at Interspeech 2026.

Martijn Wieling, Jack Grieve, Gosse Bouma, Josef Fruehwald, John Coleman, and Mark Liberman. 2016. Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, 6(2):199–234.

Mario Zusag, Laurin Wagner, and Bernhard Thallinger. 2024. CrisperWhisper: Accurate timestamps on verbatim speech transcriptions. In *Proceedings of Interspeech 2024*.

Segment-Level Register Analysis of the Multilingual Web

Erik Henriksson[1], Alireza Razzaghi[1], Tomasz Galica[1], Jiaqi Feng Guo[1], Anastasia Zhiling Dong[2], Urho Kunnas[1], Erofili Psaltaki[1], Tuomas Lundberg[1], Antti Kanner[1], Veronika Laippala[1]
[1]University of Turku, [2]University of Trier
Correspondence: erik.henriksson@utu.fi

Abstract

Web registers, the situationally defined varieties of online text, are almost always annotated at document level. A document, however, may contain multiple such varieties. We label 80,000 documents from the HPLT web corpus at segment level in seven languages using the open-weight Gemma 4, reaching Cohen's κ=0.73 against 1,878 human-annotated segments, comparable to previously reported document-level agreement. We find that most documents have two or more registers, and that they recur in similar positions and combinations across the studied languages. Yet on average, a single register accounts for 73% of the labeled text, and a quarter to a third of segments receive no label. Document-level labels are therefore mostly correct but incomplete. Multidimensional analysis shows that the segment-level labels correspond to linguistic differences. In all seven languages, registers are ordered the same way, from informational to involved text.

1 Introduction

How people communicate depends on the situation: news, recipes, encyclopedias and song lyrics each call for a different kind of language. These differences span all linguistic levels from vocabulary and morphosyntax to the organization of discourse, and have been documented across modalities, historical varieties, and languages (Pescuma et al., 2023). Variation of this kind is studied under the term *register variation*, where a register is a variety of text marked by the linguistic features typical of its situation of use (Halliday, 1978; Biber and Conrad, 2019; Matthiessen, 2019; Biber and Egbert, 2023). Register studies aim to describe empirically how the situation of use relates to linguistic form.

The web has been used as a language corpus since the early 2000s (Kilgarriff and Grefenstette, 2003; Baroni et al., 2009), offering textual variation far beyond traditional corpora (Biber and Egbert, 2018). In most previous work, web registers are assigned to whole documents (Kuzman and Ljubešić, 2025). Yet a document may consist of situationally distinct units (Egbert et al., 2024; Egbert and Gracheva, 2022): a recipe page, for example, may combine instructions, advertising, reader comments and website terms, as well as titles and navigation that have no register (Figure 1). A single label for such a document does not distinguish between these units, which may in part explain problems reported in web register research (Biber and Egbert, 2023), including fuzzy categories (Biber et al., 2020; Henriksson et al., 2024), low inter-annotator agreement (Egbert et al., 2015), and classification performance that does not improve with larger models (Henriksson et al., 2025b).

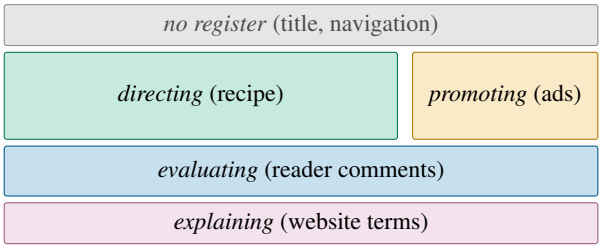


Figure 1: Schematic of register mixing on a recipe page.

Previous work has addressed these issues in two ways. The first is to let annotators assign many register labels to a single document (Laippala et al., 2023; Skantsi and Laippala, 2023). Such labels, however, only show whether a document is mixed, not whether one passage mixes two registers or two passages each have a different one. The second is to move below the document. Henriksson et al. (2025a) split English web documents recursively at the points where a register classifier predicted different registers on each side, reporting a cleaner register signal for the resulting segments. Henriksson et al. (2026) labeled paragraphs from the High Performance Language Technologies (HPLT) English web dataset (Burchell et al., 2025) with a proprietary LLM, finding that agreement is higher for segments than for documents and that over 80%

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 200–213
October 29, 2026 ©2026 Association for Computational Linguistics

of documents contain more than one register.

In this paper, we apply segment-level register analysis to seven languages, using Gemma 4 (Gemma Team et al., 2026), an open-weight local LLM, and the HPLT 4.0 corpus (HPLT Consortium, 2026). The labels agree with 1,878 segments annotated by native speakers at Cohen's κ=0.73 on average. We find that most web documents mix registers in positions and combinations that recur across languages, but that a single register usually covers most of a document's text. The segment-level labels also correspond to linguistic differences. We release our code and annotated data at `https://github.com/TurkuNLP/multilingual-web-register-segments`.

2 Background

2.1 Web register and the unit of observation

The dominant corpus-linguistic approach to web registers labels documents using the CORE (Corpus of Online Registers of English) taxonomy, which has eight main registers and optional subregisters (Laippala et al., 2023). The approach was originally developed for analyzing English web register variation (Egbert et al., 2015) but has since been extended to other languages and to train multilingual classifiers (Laippala et al., 2019; Repo et al., 2021; Skantsi and Laippala, 2023; Henriksson et al., 2025b; Razzaghi et al., 2026). CORE labels have also been used in register-aware analyses of LLM-pretraining corpora (Myntti et al., 2025; Burchell et al., 2025; Galica et al., 2026).

The limits of the document as an observation unit are recognized within this tradition. Biber et al. (2020, p. 33) note that web texts can "integrate multiple communicative purposes". Biber and Egbert (2023) observe that texts within a register vary situationally and linguistically, and that some texts belong to no register at all. Egbert and Gracheva (2022) show that situational parameters behave differently depending on the granularity of the unit. The problem is thus recognized from two directions (the unit and the categories) but this work leaves open how a smaller unit could be obtained at scale.

2.2 Register segmentation

Dividing text into coherent units has a long history in NLP (Hearst, 1994, 1997; Hearst and Plaunt, 1993; Galley et al., 2003; Liu et al., 2021). In register studies, Egbert et al. (2021) developed a method for dividing transcribed conversations into discourse units, where boundaries require a shift in communicative goal. Biber et al. (2021) applied it to 2.3 million words of the British National Corpus 2014, coding each unit for communicative purpose and identifying 16 discourse types from 9 purposes.

Henriksson et al. (2025a) segment English CORE documents with a register classifier. They fine-tune ModernBERT on full documents, then split a document recursively at the sentence boundary where the classifier's predictions for the two sides are most distinct. The segments carry a clearer register signal than the documents: a classifier trained on them reaches a micro-F1 of 0.89 against 0.86, and linguistic features vary less within a register. The method needs no segment-level training data, but it requires a document-level classifier fine-tuned for the target language, and the boundaries depend on its accuracy.

In another approach, Henriksson et al. (2026) labeled the paragraphs extracted from the XML representations in the HPLT 3.0 (Oepen et al., 2025) English dataset with a proprietary LLM, and merged consecutive segments with the same label into a single span. This approach assumes that XML-derived paragraphs of text are small enough to be roughly register-homogeneous, so that boundaries fall wherever consecutive labels differ. We adopt a similar approach here, with an open-weight LLM and further differences described in Section 3.

2.3 LLMs for multilingual register annotation

Recent work has explored LLM-based register annotation (Henriksson et al., 2024, 2026), but has relied on proprietary API models and focused on English. The multilingual capabilities of LLMs for register analysis remain largely untested (though see Idahl et al., 2026, for related work). Segment-level annotation multiplies the number of labeling calls relative to full documents, making proprietary APIs costly at scale and, since endpoints and model versions change, unreliable for replication. In this work, we instead use a local open-weight model.

3 Methods

3.1 Data and sampling

We use the cleaned monolingual datasets from HPLT 4.0, where documents are distributed in shards ordered by a document quality score (Pavón et al., 2023). We take documents in descending order of quality score until we have a pool of 10 million per language. From this pool we draw a ran-

dom subsample of 10,000 documents per language. Finally, we draw 250 segments per language for human annotation, sampling over all segments of these documents. Every segment in these documents has the same probability of being sampled.

We target seven languages, chosen to vary writing system, family and corpus size: English, Finnish, Polish, Greek, Persian, Swedish and Mandarin Chinese. HPLT distributes Chinese as two separate collections by script. We sample from both, giving eight samples for seven languages.

3.2 Segments

We take the segments provided by HPLT as our labeling unit. A document's segments are the newline-separated lines of its `text` field. HPLT describes them as "paragraph-like text passages" whose boundaries come from the markup of the original document (HPLT Consortium, 2026): text is extracted with Trafilatura (Barbaresi, 2021), which outputs one line per block-level element, so that each paragraph, heading, list item, block quote, and code block begins a new segment.

In the register analyses reported below, we merge consecutive segments with identical annotations across all fields (for the fields, see Section 3.3), and these merged spans are our unit of analysis ("register segments"). Following Henriksson et al. (2026), we treat segments with no register as transparent for the merge, so that a heading or label does not split a continuous stretch of the same register. Registerless segments are excluded from the merged span but kept as separate units in the analyses.

3.3 Annotation scheme

We annotate each segment along three dimensions, summarized in Table 1: *medium*, how the text was produced; *turn*, whether others can reply; and *purpose*, the communicative act the segment performs. The dimensions are independent, except that segments labeled `cannot_rate` for medium (headings, labels, boilerplate, and other similar text) receive no turn or purpose. Segments are annotated in context: a segment that is fragmentary on its own but continues a passage is rated as part of that passage. The same codebook is given to both the human annotators and to the LLM.

To relate our annotations to previous work, we map each segment to a single CORE main register (Figure 2): *transcribed* segments become *Spoken*, the remaining *dialogic* ones *Interactive discussion*, and all others are mapped by purpose, with *explaining* and *supporting* collapsing into one register and *evaluating* and *expressing* into another. Although CORE registers were originally defined for full documents (e.g. *Narrative* covers blogs, news articles, and stories), the situational parameters underlying them (Egbert et al., 2015) are mostly comparable to those used here. The exception is *creating* (literary or artistic text), which we map to the more restrictive *Lyrical* CORE register (songs and poems).

Medium — *how the text was produced*	
written	composed to be read
transcribed	a record of words actually spoken
cannot_rate	not self-standing communication: headings, labels, boilerplate, metadata, fragments, machine-generated or garbled text
Turn — *whether others can reply*	
dialogic	a real exchange where others can reply
monologic	the audience cannot reply
Purpose — *the communicative act performed*	
explaining	conveying how something is or works
supporting	defending a specific claim with evidence
recounting	relating particular real events
directing	telling the reader (how) to do something
evaluating	arguing a position, backed by reasons
expressing	a feeling or reaction for its own sake
promoting	marketing own product or organisation
creating	literary or artistic text as the main content

Table 1: The annotation scheme. *Turn* and *purpose* are left empty when the medium is `cannot_rate`. Glosses are abridged; Appendix A gives the codebook in full.

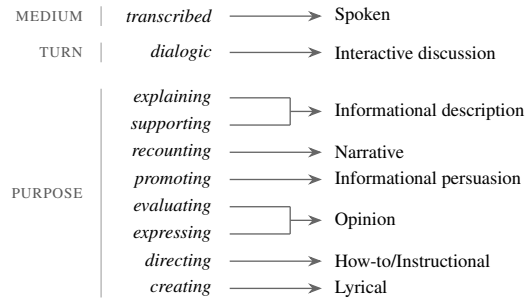


Figure 2: Mapping of segment annotations to CORE main registers.

3.4 Human annotation

For each language, 148–250 sampled segments were annotated by one native speaker of that language, using a browser-based annotation tool. Annotators followed the same codebook as the LLM (Appendix A) and saw the same context: the target segment and its ±5 neighboring segments, the first five segments of the document, and the page URL.

The annotations were also checked for codebook consistency. We used Claude Opus 5 (Anthropic,

2026) to flag labels that were in clear conflict with the guidelines, and one author reviewed the flagged cases, using machine translation as necessary, and fixed obvious errors. This post-correction step reduces the impact of annotator errors on the figures.

3.5 LLM annotation

We annotate with Gemma 4 31B Instruct (Gemma Team et al., 2026), a dense open-weight instruction-tuned model, served with vLLM (Kwon et al., 2023) on two AMD MI250 GCDs of the LUMI supercomputer. Decoding is greedy (temperature 0) and constrained to a JSON schema with XGrammar (Dong et al., 2024), so the model can only output the six fields of the scheme.

The prompt gives the model the codebook, the page URL, and a context block in which the target segment is marked and its neighbors are shown (Appendix B). Before outputting the labels, the model writes a one-sentence rationale for its decision. The schema places this `reasoning` field first, and because the grammar generates the fields in order, the model must state its reason before labeling.

To check whether the open-weight model is competitive on this task, we repeat the annotation with Claude Opus 5 and compare the two sets of labels. The setup is identical, except that the constrained JSON is obtained through forced tool use, the equivalent mechanism in the Anthropic API.

3.6 Evaluation

We score the model against the human annotation per dimension. *Medium* and the CORE-mapped register are scored over all segments, since both are defined for every segment. *Turn* and *purpose* are scored over the segments that both sides judged ratable, because they are undefined where medium is *cannot_rate*. The joint *all* column requires the model to match the human on all three dimensions. We report F1 and Cohen's κ (Cohen, 1960).

3.7 Multi-dimensional analysis

To test whether the labels correspond to linguistic differences, we run a multi-dimensional analysis (Biber, 1988) over the register segments. Segments of at least 50 words are parsed with Stanza (Qi et al., 2020) and described by rates of universal part-of-speech (UPOS) tags, universal dependency relations (de Marneffe et al., 2021) and punctuation types per 1,000 content tokens, plus eight structural measures (sentence and word length, dependency distance, tree depth, long and head-final dependencies, lexical density, type–token ratio). Features are standardized within each sample, so that dimensions capture within-language covariation rather than cross-language contrast. After filtering (Appendix C), 47 features over 425k segments are factored by iterated principal axis factoring with promax rotation; a segment's score on a dimension sums its standardized features loading at $|\lambda| \geq 0.35$.

	Cohen's κ		F1	
Field	Gemma 4	Opus 5	Gemma 4	Opus 5
Medium	0.79	0.80	0.90	0.91
Turn	0.86	0.78	0.97	0.96
Purpose	0.70	0.68	0.74	0.72
CORE label	0.73	0.72	0.78	0.78

Table 2: Agreement between each LLM annotator and the human annotation, averaged over the eight language samples (n=1,878 segments). The *turn* κ excludes both Chinese samples, where too few segments are dialogic.

4 Results and discussion

We first evaluate the LLM-generated labels against the human annotation, then use the labels to describe the corpus, looking at how registers combine and how they are distributed within documents. Finally, we examine how the segment-level registers differ linguistically.

4.1 Annotation quality

Gemma 4 reaches κ=0.73 on the CORE-mapped label on average, falling between 0.72 and 0.74 in every sample (F1 0.76–0.79). This resembles the human–human agreement in the CORE tradition: 83% for the FinCORE main registers (Skantsi and Laippala, 2023) and 78–84% F1 for FreCORE and SweCORE (Repo et al., 2021), against Fleiss' κ=0.47 for English CORE (Egbert et al., 2015). At the segment level, Henriksson et al. (2026) report κ=0.76 for English. The labels thus appear to be close to the ceiling the task allows.

Claude Opus 5, a proprietary frontier model, does not do better on this task (κ=0.72), and the two models agree with each other on 81.2% of segments, more often than either agrees with the annotators (78.2% and 77.8%; p=0.003). This suggests the disagreements are a property of the task and the codebook rather than of the chosen model. Below, we report the results for Gemma 4, the model we use for the large-scale annotation (Section 4.2).

The model and the annotator disagree on at least one dimension in 26% of the segments. Disagree-

	model									
human	expl	supp	reco	dire	eval	expr	prom	crea	n.r.	n
expl	**308**	5	11	7	9	–	8	1	11	360
supp	16	**10**	6	–	20	–	–	–	1	53
reco	15	–	**181**	1	6	2	1	–	1	207
dire	13	–	3	**116**	5	–	2	–	6	145
eval	16	1	5	8	**120**	1	2	–	–	153
expr	12	–	24	4	36	**39**	1	–	3	119
prom	29	–	5	5	9	–	**83**	–	3	134
crea	1	–	8	1	2	1	–	**48**	2	63
n.r.	66	1	10	19	9	10	8	5	**516**	644
n	476	17	253	161	216	53	105	54	543	1878

Table 3: Confusion between the human annotation and Gemma 4 31B on *purpose*, with the segments carrying no register as a ninth category, pooled over the eight samples (n=1,878 segments). "n.r." = `cannot_rate`.

ments fall mainly on *purpose* (63%) and ratability (32%); turn and the *written/transcribed* distinction account for the remaining 5%. Table 3 shows the first two together. Ratability errors are asymmetric: the model assigns `cannot_rate` less often than the annotator (precision 0.95, recall 0.80). Recall is 0.89 for segments under 40 characters but 0.62 for those over 150, so the missed cases are most often long texts that the model judges ratable.

Within *purpose*, two of the eight categories perform poorly. The model recovers 19% of the annotator's *supporting* segments and 33% of the *expressing* ones (F1 0.29 and 0.45), against 0.65–0.82 for the other six. The scheme separates both from a close neighbour by degree: *supporting* is separated from *explaining* by whether the text is defending a claim or just stating it, and *expressing* from *evaluating* by whether the stance is argued for. Both are absorbed mainly by *evaluating*: 20 of the 53 *supporting* segments and 36 of the 119 *expressing* ones are labelled *evaluating*, which the model uses for 216 segments against the annotator's 153. Many of these cases are genuinely ambiguous. In the comment below, the annotator read the target as a bare reaction (*expressing*) and the model as an argued position (*evaluating*):

> *[Context −1]* … My blog is the one place — besides with my oldest friends — where I am the closest to who I am at my core.
>
> **[Target]** Finally, there are many, many ways to monetize blogs beyond ads, brands, and reviews. Good for the people that make it work, but we can all expand.
>
> *[Model rationale]* "The author argues that bloggers should expand their monetization methods beyond ads and brands."

Both readings follow the codebook (Appendix A):

Language	Segments	Mchar	no reg.	tra.	dia.
English	150,111	158.3	26.6	4.3	6.3
Finnish	142,626	76.2	31.0	1.6	10.0
Polish	124,585	75.1	35.4	3.5	6.1
Greek	139,457	84.7	30.2	5.2	6.6
Persian	167,024	91.4	32.6	7.1	6.0
Swedish	147,955	85.3	31.1	2.3	5.3
Chinese (S)	113,456	47.2	27.7	4.6	5.0
Chinese (T)	93,027	23.1	29.8	1.6	1.7
Total	1,078,241	641.4			

Table 4: The labeled corpus: 10,000 documents per sample. A segment is a merged span (Section 3.2), with `cannot_rate` segments transparent to the merge. *no reg.* = `cannot_rate`; *tra.* = transcribed and *dia.* = dialogic.

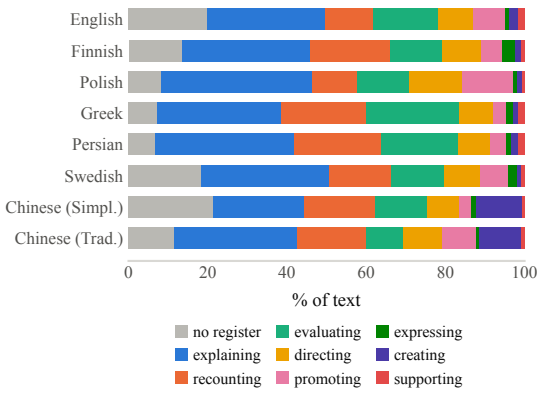


Figure 3: Composition of each sample. Each bar is all of that sample's text: in grey the share of segments labelled `cannot_rate`, the rest divided by *purpose*.

the segment recommends a course of action but gives no reasons for it. The same confusions occur in all languages, but in different amounts: per-sample agreement on *purpose* ranges from κ=0.61 to 0.75, with the three lowest samples containing the most *supporting* and *expressing* segments. Under the CORE mapping, which merges *explaining* with *supporting* and *evaluating* with *expressing*, agreement is more stable (0.72–0.74).

We conclude that the disagreement is concentrated in two category boundaries, and not in any particular language. Once *supporting* and *expressing* are merged with their neighbors under the CORE mapping, agreement is nearly identical across the eight samples.

4.2 Register distribution across languages

The eight samples have similar register compositions (Table 4, Figure 3). A segment here is a merged span of consecutive identical annotations (see Section 3.2); the 2.1 million annotated HPLT segments are merged into the 1.1 million regis-

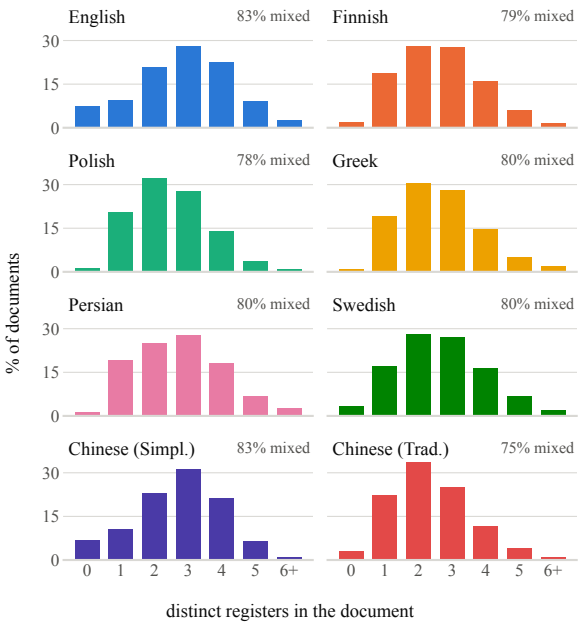


Figure 4: Distinct CORE registers per document. The scheme has eight registers, but counts of >=6 are pooled, as together they hold 0.7–2.6% of documents.

ter segments reported in the table (9–17 per document). Between 26.6% and 35.4% lack register (including headings, labels, navigation, metadata and fragments). Measured in text (Figure 3) they are smaller, holding 6.8% (Persian) to 21.4% (Chinese Simpl.). Transcribed text accounts for 1.6–7.1% of all text and dialogic text for 1.7–10.0%.

Figure 3 shows the distribution of purposes in the labeled segments. *Explaining* is the largest in every sample, from 29.3% (Chinese Simpl.) to 41.6% (Polish) of ratable text, and *explaining*, *recounting*, *evaluating* and *directing* together account for 76.5–91.6%. *Creating* is the most different between samples, taking 14.7–11.9% of ratable text in the two Chinese samples against at most 2.7% in any other.

4.3 Register mixing within documents

Most documents contain more than one register, in all eight samples. Figure 4 counts the distinct CORE registers (Figure 2) in a document. We count CORE registers here rather than the three original dimensions to make the figure comparable with previous work (e.g. Biber and Egbert, 2018; Henriksson et al., 2025b). `cannot_rate` is the absence of a register, so a document with only `cannot_rates` is counted as having 0 registers. Between 75.0% (Chinese Trad.) and 83.1% (English) of documents have two or more, most commonly two or three, with a mean of 2.4 to 2.9. Only 9.5% (English) to 22.1% (Chinese Trad.) are register-homogeneous, and 0.9% (Greek) to 7.4% (English) lack register. For English, the figure agrees with the mixing rate Henriksson et al. (2026) report at segment level.

4.4 Positional distribution of registers

To see where registers fall within a document, we count characters in each category by relative position in 128 equal-width bins (Figure 5). Each category is normalized over its own bins and expressed as a ratio to an even spread, so the curves show positional shape. We plot *dialogic* and *transcribed* but not their opposites, since *monologic* (93%) and *written* (95%) reproduce the corpus average. *Supporting* is omitted (1.3% of the text).

The patterns are clear and nearly identical across the eight samples. Registerless text concentrates at page edges, reaching 2.3 times an even spread in the first bin and 1.8 in the last, likely reflecting titles and navigation at the top and boilerplate at the end. *Recounting* tends to appear early, peaking in the first quarter and falling to 0.7 by the end, while *explaining* is spread almost evenly, varying only between 1.1 and 0.8. *Dialogic* and *expressing* instead concentrate at the end: *dialogic* rises from 0.5 to 1.3 across the document, and *expressing* dips to 0.7 at the quarter mark before rising to 2.6. Both point to document-final comments, where opinionated text is common.

4.5 Register dominance

Section 4.3 above reports register *counts* per document (Figure 4). Weighting by *content* shows a different picture: it is dominated by a single register. Figure 6 shows, averaged over documents, the share of register-bearing text held by the largest CORE register, the 2nd largest, and so on. Across the eight samples, the largest holds 69.0–74.6%, the second 18.3–21.4% and the third 5.2–7.4%.

The profiles are similar across languages and slightly more concentrated than chance. We compare each page against a random division of its text among the registers it contains: for k registers, the expected largest share is $(1/k)\sum_{i=1}^{k} 1/i$, or 75% for two registers and 61% for three. The observed share exceeds this baseline in all eight samples, by 1.6–5.5 points. The comparison holds the number of registers fixed, so it tests how evenly a page's text is divided among its registers. Although most pages are multi-register, they are usually organized around one dominant register.

The text outside the dominant register follows

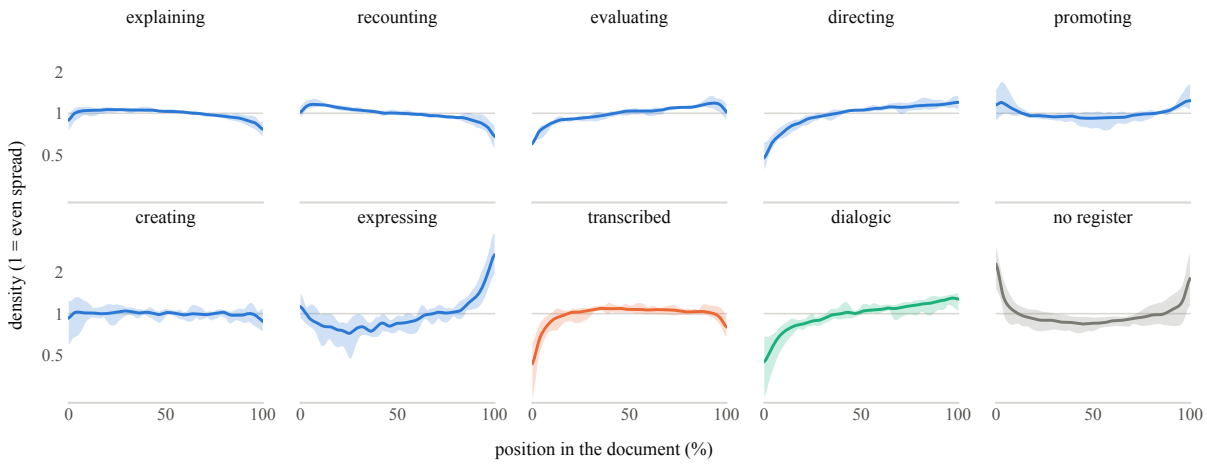


Figure 5: Positional distribution of registers within documents, as a ratio to an even spread (1, logarithmic axis) over 128 bins, Gaussian-smoothed ($\sigma = 2$). Lines are means over the eight samples, bands show the full range.

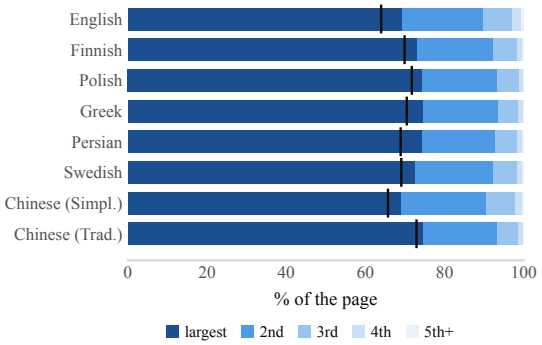


Figure 6: Register profiles of documents. Mean share of labeled text held by the largest register, the second largest, and so on. Black marks give the expected share of the largest register under random division of a page into the number of registers it actually contains, computed per sample. `cannot_rate` segments are excluded.

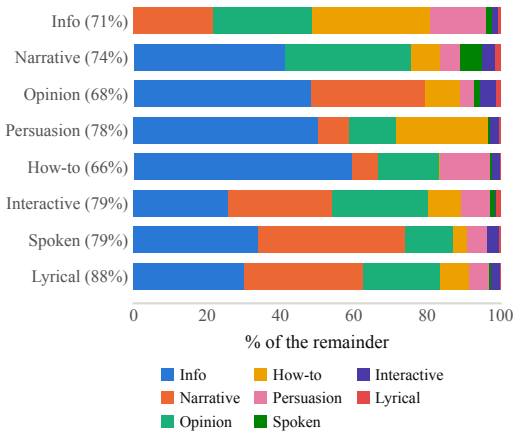


Figure 7: The mean composition of the text on the page accompanying the main register's share (shown in brackets). Shares are of the remainder of the page, averaged over documents and pooled across languages.

regular patterns too. Figure 7 groups documents by their dominant register and shows the composition of the rest of the page. The share of pages with text outside the dominant register ranges from 64% to 89%, depending on which register dominates. On a page dominated by Informational description, the rest is mostly How-to (32%), Opinion (27%) and Narrative (22%). On a How-to page, it is mostly Informational description (60%).

The figures above partly reflect how common each register is overall. Figure 8 therefore compares each register's share of a remainder against its share of remainders elsewhere (1 = no difference). Informational description, How-to and Informational persuasion co-occur more often than expected (1.1–1.6), and the last two co-occur with Narrative, Spoken and Lyrical *less* often than expected (0.2–0.4). Narrative forms a second group with Opinion, Interactive discussion, Lyrical and Spoken (1.3–2.5, and 1.3–1.9 in the reverse direction). The association is strongest with Spoken, but these four do not co-occur consistently with each other. The two groups correspond to the Informational and Involved poles reported in multi-dimensional analyses (Biber and Conrad, 2019).

4.6 Linguistic dimensions

The segment-level labels correspond to linguistic differences. We show this with a multi-dimensional analysis of the register segments (Biber, 1988). The features are taken from Stanza parses (see Section 3.7) and the dimensions are derived from their co-variation. Four dimensions were extracted, account-

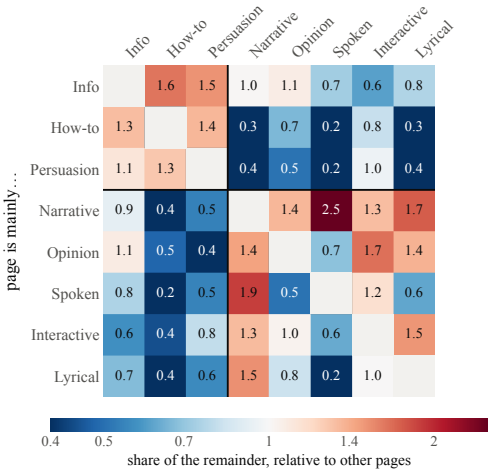


Figure 8: Remainder composition (rows: dominant register; columns: registers in the remainder). Each cell gives a register's share of that remainder. Values are relative to the same register's share of remainders on pages with a different dominant register, so 1 means no difference. The rule separates the informational registers from the narrative and interpersonal ones.

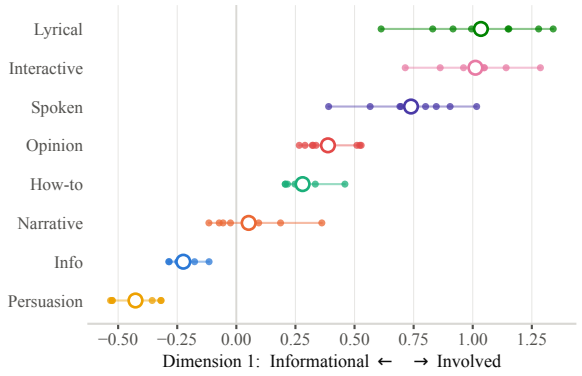


Figure 9: Register means on Dimension 1, in standard deviations from the corpus mean. Small marks give the eight sample means and the line their range; the open circle is the mean over samples.

Positive	verbs (.78), clause markers (.71), pronouns (.71), subordinators (.70), auxiliaries (.65), adverbial modifiers (.64), subjects (.60), adverbs (.60), auxiliary dependents (.54), clausal complements (.53), adverbial clauses (.52), open clausal complements (.46), copulas (.39), objects (.38)
Negative	lexical density (-.66), nominal modifiers (-.65), nouns (-.60), word length (-.55), adjectival modifiers (-.37)

Table 5: Features loading at ≥ 0.35 on Dimension 1.

ing for 33.3% of the variance. We report the first, which accounts for 15.4%; the remaining three account for 5.5–6.6% each.

Dimension 1 corresponds to the opposition between *Informational* and *Involved* production that is reported throughout the multi-dimensional literature (Biber, 1988, 1995; Biber and Conrad, 2019). Its positive pole consists of verbs, pronouns, auxiliaries, adverbs, subordinators, subjects and subordinate clauses of three kinds, and its negative pole of lexical density, nominal modifiers, nouns and word length (Table 5). Although no register information enters the analysis, register accounts for 18.3% of the variance in the resulting scores.

Figure 9 shows each register's mean score in each of the eight samples, together with the pooled mean. The involved pole is occupied by Lyrical and Interactive discussion (both $+1.0$), followed by Spoken ($+0.74$); the informational pole by Informational persuasion (-0.43) and Informational description (-0.22), with Opinion, How-to and Narrative between them. The ordering is largely consistent across the eight samples: Spearman's ρ between each sample's ordering and the pooled one ranges from 0.88 to 1.00, and every register at either pole stays there, Lyrical, Interactive discussion and Spoken positive in all eight samples and the two informational registers negative. How strongly register predicts the score is less even, from 12.8% of the dimension's variance in Chinese (Trad.) to 25.9% in Finnish. The same Informational–Involved opposition was inferred in Section 4.5 from the registers that co-occur within a document; it is obtained here from the linguistic features directly.

Segments labeled `cannot_rate`, omitted from the figure, fall below the informational pole (-0.83; -1.61 in Persian), indicating that the text carrying no register is a distinct, densely nominal variety. This holds in six of the eight samples; in both Chinese samples they are just above Informational persuasion. However, only 11.5% of registerless segments reach the 50-word floor (Section 3.7) against 59–72% for the eight registers, so this analysis characterizes only the longer minority.

5 Conclusion

We annotated the register of 80,000 web documents at segment level in seven languages and four scripts, using Gemma 4 and validating the labels against native-speaker annotation. Agreement reaches κ=0.73 for the CORE mapping and varies little across the eight samples (0.72–0.74). Claude Opus 5 does not perform better, and the two models agree with one another more often than ei-

ther agrees with the annotators. The disagreements are therefore not specific to the model we use.

Two distinctions in the scheme, between *explaining* and *supporting* and between *evaluating* and *expressing*, are not applied reliably at segment level, and merging them under the CORE mapping gives stable agreement across the samples.

In general, the language samples are very similar. Two or more registers occur in 75–83% of documents; *explaining*, *recounting*, *evaluating* and *directing* account for most of the text; the registers occupy similar positions within the document, with expressing and dialogic text concentrated at the end; and the multi-dimensional analysis yields the same *Informational–Involved* opposition, ordering the registers in the same way, in all eight samples.

Although most documents mix registers, a single one covers 73% of the register-bearing text in an average document, and the remainder follows regular patterns, with informational registers grouping apart from narrative and interpersonal ones.

Therefore, we conclude that document-level labeling, typical of previous work, describes only part of most pages. It identifies the dominant register, missing on average a quarter of the register-bearing text and all that lack one. For register-aware curation of LLM corpora (Myntti et al., 2025; Burchell et al., 2025; Galica et al., 2026), segment-level labels allow for finer-grained data selection.

6 Limitations

With one annotator per language, we report model–human rather than human–human agreement, and comparisons to published IAA scores are indicative, as they come from different studies and label sets. The human labels were post-corrected with model assistance, which may inflate the agreement.

We use HPLT's segmentation, which is not linguistically motivated: boundaries may not fall where the register changes, and segments can have internal shifts. Merging same-label segments corrects the first problem, but a mixed segment still receives a single label; locating these shifts, with multi-label assignment, is left for future work. The languages were chosen by annotator availability and are well resourced in HPLT, and we use only high-quality text. Other languages and lower-scoring text might yield different results.

The multi-dimensional analysis omits morphology for cross-language comparability, and depends on parser accuracy that varies by language. Finally, we use a single model and a prompt developed informally rather than tuned against a development set; future work could optimize prompts systematically, e.g., with DSPy (Khattab et al., 2023).

Ethics statement

Our data comes from HPLT 4.0, which is built from web crawls. Web text of this kind contains personal information. We use the corpus as distributed and have not removed such content, since that would alter the distribution of text we set out to describe. HPLT drops documents from domains whose `robots.txt` disallows the crawlers it uses, flags any personal information detected, and gives rights holders an address for take-down requests, acting on them in subsequent releases (HPLT Consortium, 2026). Our own annotations describe the communicative situation of passages of text; we add no personal information to what HPLT already publishes.

Acknowledgments

This work was supported by the Research Council of Finland through the FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union NextGenerationEU instrument), the project "Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora" (project 362459), and the Centre of Excellence "Human Diversity through Contacts" (project 374223). Further support came from the OpenEuroLLM and LLMs4EU projects under Digital Europe Programme grant agreements No 101195233 and 101198470, and from the European Union's Horizon Europe research and innovation programme under Marie Skłodowska-Curie grant agreement No 101177564 (HAIF), co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. We also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Anthropic. 2026. Claude opus 5.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Dis-

covery and Extraction. In *Proceedings of the Joint Conference 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. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Douglas Biber. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2023. What is a register?: Accounting for linguistic and situational variation within–and outside of–textual varieties. *Register Studies*, 5(1):1–22.

Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.

Douglas Biber, Jesse Egbert, Daniel Keller, and Stacey Wizner. 2021. Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics*, 171:20–35.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 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*, volume 1: Long papers, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2024. XGrammar: Flexible and efficient structured generation engine for large language models. *arXiv preprint arXiv:2411.15100*.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Jesse Egbert, Douglas Biber, Daniel Keller, and Marianna Gracheva. 2024. Register and the dual nature of functional correspondence: accounting for text-linguistic variation between registers, within registers, and without registers. *Corpus Linguistics and Linguistic Theory*, 20(3):505–538.

Jesse Egbert and Marianna Gracheva. 2022. Linguistic variation within registers: Granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory*, 19.

Jesse Egbert, Stacey Wizner, Daniel Keller, Douglas Biber, Tony McEnery, and Paul Baker. 2021. Identifying and describing functional discourse units in the bnc spoken 2014. *Text & Talk*, 41(5-6):715–737.

Tomasz Galica, Amanda Myntti, Erik Henriksson, Fedor Vitiugin, Sampo Pyysalo, and Veronika Laippala. 2026. Register-based data selection for multilingual llms. Accepted at EMNLP 2026.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Michael Alexander Kirkwood Halliday. 1978. *Language as Social Semiotic*. Arnold, London.

Marti A. Hearst. 1994. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Marti A. Hearst and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in*

Information Retrieval, SIGIR '93, page 59–68, New York, NY, USA. Association for Computing Machinery.

Erik Henriksson, Saara Hellström, and Veronika Laippala. 2025a. Analyzing register variation in web texts through automatic segmentation. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 7–19, Albuquerque, USA. Association for Computational Linguistics.

Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, and Veronika Laippala. 2024. From discrete to continuous classes: A situational analysis of multilingual web registers with LLM annotations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 308–318, Miami, USA. Association for Computational Linguistics.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2025b. Automatic register identification for the open web using multilingual deep learning. *Preprint*, arXiv:2406.19892.

Erik Henriksson, Alireza Razzaghi, Tuomas Lundberg, Antti Kanner, and Veronika Laippala. 2026. Register mixing is the norm on the web. In *Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities*, pages 138–149, San Diego, USA. Association for Computational Linguistics.

HPLT Consortium. 2026. HPLT 4.0 Monolingual Data Pool. `https://hplt-project.org/datasets/v4.0`. Version 4.0, accessed 2026-08-05.

Maximilian Idahl, Benedikt Droste, Björn Plüster, and Jan Philipp Harries. 2026. propella-1: Multi-property document annotation for llm data curation at scale. *Preprint*, arXiv:2602.12414.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *Preprint*, arXiv:2310.03714.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

Taja Kuzman and Nikola Ljubešić. 2025. Automatic genre identification: A survey. *Language Resources and Evaluation*, 59(1):537–570.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.

Yang Liu, Chenguang Zhu, and Michael Zeng. 2021. End-to-end segmentation-based news summarization. *Preprint*, arXiv:2110.07850.

Christian MIM Matthiessen. 2019. Register in systemic functional linguistics. *Register Studies*, 1(1):10–41.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. Register always matters: Analysis of LLM pretraining data through the lens of language variation. In *Conference on Language Modeling (COLM)*, Montreal, Canada.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, et al. 2025. Hplt 3.0: Very large-scale multilingual resources for llms and mt. mono-and bi-lingual data, multilingual evaluation, and pre-trained models. *arXiv preprint arXiv:2511.01066*.

Pablo Pavón et al. 2023. Web docs scorer. `https://github.com/pablop16n/web-docs-scorer`. GitHub repository.

Valentina N. Pescuma, Dina Serova, Julia Lukassek, Antje Sauermann, Roland Schäfer, Aria Adli, Felix Bildhauer, Markus Egg, Kristina Hülk, Aine Ito, Stefanie Jannedy, Valia Kordoni, Milena Kuehnast, Silvia Kutscher, Robert Lange, Nico Lehmann, Mingya Liu, Beate Lütke, Katja Maquate, and 24 others. 2023. Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods. *Frontiers in Psychology*, Volume 13 - 2022.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alireza Razzaghi, Erik Henriksson, and Veronika Laipalla. 2026. ParsCORE: The Persian corpus of online registers. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 60–73, Rabat, Morocco. Association for Computational Linguistics.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas

Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. *arXiv preprint arXiv:2102.07396*.

Valtteri Skantsi and Veronika Laippala. 2023. Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics*, 48(1):1–31.

A Annotation codebook

The following is the codebook given to the human annotators and embedded verbatim in the model prompt. Only its typesetting has been adapted here.

Order of decisions. Start with medium. If it is `cannot_rate`, leave the other values empty. Web pages contain a great deal of text that is not self-standing communication; expect roughly a third of segments to be `cannot_rate`. Do not treat *written* as a fallback for anything you cannot otherwise place — decide `cannot_rate` on its own merits first.

`mode_medium` (always pick one). A segment is `cannot_rate` when it does not work as a self-standing piece of human communication: titles, headings, labels, menus, link lists, boilerplate, metadata, captions, dates, bare fragments, machine-generated text, and garbled text. A list item qualifies only if it is a complete statement on its own.

Headings are headings whatever their grammar. A segment that names what the following segments are about is a heading even when it is a full sentence or a question. The test is positional, not grammatical: the next segment starts the exposition rather than continuing the target's sentence, and the target announces a topic instead of saying anything about it. Question-form headings ("Why is X worth having?", "How do I do Y safely?") are common on generated articles and are `cannot_rate`.

Judge the segment on its own text, not on the page around it. A segment is `cannot_rate` only when its own words do not form coherent communication. A segment that reads as a coherent statement stays ratable even when the surrounding page is machine-generated, even when it is logically unrelated to the segments around it, and even when it contains typos or a wrong word. Spun or machine-translated text is `cannot_rate` only where the segment itself is word salad: unrelated words, product names, or other-language fragments spliced together inside the one segment; implausible synonyms that leave the segment itself meaningless ("Bond release of Forex Megadroid", "the too much to handle clutter"); the search keyphrase stuffed into the segment; or leftover spinner markup ("{Reciprocal}") and words run together ("andyour"). A blurb on a directory or listing page is ratable when the blurb itself is a coherent statement, even when its neighbours are unrelated entries.

Judge this in context, not in isolation. A segment may look like a fragment on its own but read as part of a continuous passage with the segments before and after it (for example, a single sentence split across segments). When it is coherent in context, rate it as part of that passage. This is the test for list items too: a list item is ratable when it reads as a complete statement, whether on its own or together with its surrounding context. Dates, headings, and labels stay `cannot_rate`.

For anything that is ratable, pick how it was made. *Transcribed*: a record of words actually spoken aloud — transcripts, interviews, speeches, captions or subtitles, or text carrying speaker labels or turn markers; use this only when the text is clearly a transcription of speech. *Written*: everything else, text composed to be read; casual or chatty tone does not make it transcribed (comments, posts, messages, and emails are all written).

`mode_turn` (skip if `cannot_rate`). *Dialogic* only when the text is part of a real exchange where others can reply: forums, comments, social posts, chats, interviews. *Monologic* otherwise: one author addressing an audience that cannot reply in place (articles, guides, FAQs). This is the default. Dialogue inside a fictional or scripted work is still monologic, since it has one author and is not a real interaction.

`field_purpose` (skip if `cannot_rate`).

- *explaining* — conveying how something is or works so the reader understands it: concepts, definitions, mechanisms, background. The information is presented as established fact, not as a claim the author is out to prove. E.g. an encyclopedia-style paragraph on how a process works.
- *supporting* — establishing a specific factual claim by offering evidence for it: data, study results, sources, or firsthand findings presented as proof. Typical of research and evidence-based writing. If the text just presents accepted information without defending a contested claim, it is explaining, not supporting.

- *recounting* — relating specific real events or experiences presented as having actually happened: a trip report, news account, incident writeup, or personal anecdote. May be a single event and need not be in time order — what matters is that these are particular real happenings. Not general facts (explaining), not a summary of fiction (a plot recap is explaining, not recounting), and not events still to come.
- *directing* — telling the reader to do something, or how to do it: instructions, steps, recipes, how-to guidance, advice on what to choose. Impersonal advice still directs ("do X", "one should do X"), but a statement that something is important or beneficial, or a document's own rules and legal terms, is explaining — directing tells the reader to act.
- *evaluating* — arguing a position: a judgment, recommendation, or claim about what is good or bad or what should be done, backed by reasons. Covers reviews, opinion pieces, and advocacy — anything that makes a case the reader could agree or disagree with.
- *expressing* — voicing a feeling, reaction, or stance for its own sake, without arguing for it. The point is how the author feels, not a case they make. E.g. a comment that simply says the author loves or hates something.
- *promoting* — the author marketing their own product, service, or organization.
- *creating* — literary or artistic text as the main content: poetry, lyrics, scripture, fiction, crafted narrative prose. Story form alone is not recounting; if it reads as fiction or literature, it is creating.

Prefer the simplest purpose that fits; only reach for a richer label when its defining feature is actually present. Use *evaluating* only when the author argues a position (not for merely opinionated or topical text), *supporting* only when evidence is offered for a claim (not for plain description), and *recounting* only for specific real events (not general facts, and not summaries of fiction).

When two are close. *Evaluating* vs. *expressing*: reasons given = evaluating; bare reaction = expressing. Ask whether the reader could argue with the reasoning or only with the taste. Opinion pieces and advocacy that give reasons are evaluating, even when passionate. *Explaining* vs. *supporting*: neutral, accepted information you are meant to understand = explaining (an encyclopedia entry); a specific claim defended with evidence so you will believe it = supporting (a study's results). Facts alone are not supporting — ask whether the passage is proving a point or just informing. *Explaining* vs. *recounting*: general, timeless facts = explaining; specific things that actually happened = recounting. *Explaining* vs. *directing*: how something works, told so you understand it = explaining; what you should do about it = directing. Advice phrased as a fact about what is important or effective is still directing. *Explaining* vs. *promoting*: promoting is the author marketing their own product, service, or business to attract customers — a seller's own description of what they offer is promoting even when the tone is factual. Two limits keep it from over-firing: a third party describing an organisation it does not own is explaining, and text describing a product's drawbacks or harms is not promoting. *Recounting* vs. *creating*: really happened = recounting; fiction/literary = creating. *Recounting* vs. *expressing*: events for their own sake = recounting; events as setup for a present feeling or plea = expressing.

B Prompt

The user prompt is the codebook of Appendix A, followed by the output instructions below, the page URL, and the rendered context block. The system message is: *"You are an expert linguistic annotator. You label one TARGET text segment at a time along three dimensions, using the surrounding segments only as context. You respond with a single JSON object and nothing else."*

Output instructions. These state in words what the annotation tool enforces through its interface for the human annotators.

```
Output:
  - Label the TARGET segment only; use
    the surrounding segments as context.
  - Always give mode_medium. If
    mode_medium is cannot_rate, set the
    other five fields to null.
Respond with a single JSON object with
exactly these keys and nothing else:
  mode_medium, mode_turn, field_purpose
```

In the *reasoning* configuration used throughout the paper, the schema additionally carries a leading `reasoning` string and the prompt ends with: *"First fill the* `reasoning` *field with ONE short sentence (25 words max) naming the deciding cue for the target's purpose (and its medium if unclear). Then give the labels. Do not exceed one sentence."*

Context block. Segments are rendered with their offset from the target; where the local window does not reach the start of the document, the first five segments are shown separately as the document opening. The following is a real example, illustrating why segments must be judged in context: HPLT's segmentation has split a single sentence of an abstract across several segments, and the document opening consists of navigation labels.

```
URL: https://www.researchgate.net/
     profile/[anonymized]

[Document start]
  [seg 0] Network
  [seg 1] Followers
  [seg 2] Following
  [seg 3] Research
  [seg 4] Research Items (196)
  [... segments omitted ...]
[Context around the target]
  [-5] calibration and archival at the
       scale of HUC 12 watersheds
  [-4] (Hydrologic Unit Code) anywhere in
       the continental US. Our prototype
  [-3] for model-data integration focuses
       on creating tools for fast data
  [-2] storage from selected national
       databases, as well as the
  [-1] computational resources necessary
       for a dynamic, distributed
       watershed simulation. The paper
>>> [TARGET] will describe cyber-
    infrastructure tools and workflow
    that attempts to
  [+1] resolve the problem of model-data
       accessibility and scalability such
  [+2] that individuals, research teams,
       managers and educators can create a
  [+3] WVO in a desired context. Examples
       are given for the NSF-funded Shale
  [+4] Hills Critical Zone Observatory and
       the European Critical Zone
  [+5] Observatories within the SoilTrEC
       project. In the future
```

Both the human annotator and the model labelled this segment *written / monologic / explaining*; the model's rationale was "The text describes the contents and goals of a paper, which is a factual explanation of the work's scope."

C Multi-dimensional analysis: features and filtering

Parsing. All eight samples are parsed into Universal Dependencies (de Marneffe et al., 2021) with Stanza (Qi et al., 2020). We parse only segments whose segment-level language, identified with OpenLID, matches the collection's, which keeps 87.9–97.4% of them.

Feature selection. A feature enters the analysis only if it is attested in all eight samples. This excludes the entire morphological layer: the morphological feature layer `FEATS` is language-specific, and the eight treebanks share none of it; Chinese annotates 13 feature–value pairs against the 84 in Polish, and Swedish marks no verb person. We then remove features correlating above $r = 0.99$, features that are exact linear functions of the rest, and features whose squared multiple correlation falls below 0.20, leaving 47.

Factoring. Features are standardized within their own sample before the samples are pooled, so that a dimension is a pattern of within-language covariation rather than a contrast between languages. The correlation matrix is factored by iterated principal axis factoring and rotated with promax ($\kappa = 4$). Bartlett's test rejects sphericity ($p < 10^{-16}$) and the Kaiser–Meyer–Olkin measure is 0.665. A segment's score on a dimension is the sum of its standardized features loading at $|\lambda| \geq 0.35$, with negative loadings subtracted (Biber, 1988).

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

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 214–219
October 29, 2026 ©2026 Association for Computational Linguistics

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	–	–	–	–	–	–
GlotCC-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 (Raffel 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 *cho*	14	9	12	88%	6%
FW2 *mus*	10	7	5	45%	26%
GlotCC *cho*	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).

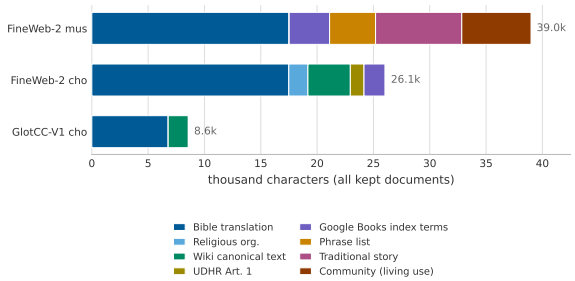


Figure 1: Genre composition (characters) of every kept Muskogean document across the three corpus subsets. Bar labels are rounded to 0.1k characters; totals in the text use exact counts.

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
cho	853	59	99.7%	27	14
mus	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 Ekvn-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 (`mvskokehymns.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	*cho*_Latn	5	5/5	5/5	0.75
Muscogee	*mus*_Latn	4	4/4	4/4	1.00
Chickasaw	*none*	2	→ cho_Latn (0.74)		
Alabama	*none*	1	→ eng_Latn (0.33)		
Koasati	*none*	1	→ rel_Latn (0.88)		
Mikasuki	*none*	1	→ bea_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 Choctaw-ish 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 worldwide, 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 reidentified 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 fastText `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.

The Crawl Gap: Where Endangered-Language Web Content Is Lost Before It Reaches a Corpus

Sharvin Goyal
Independent Researcher
sharvingoyal09@gmail.com

Abstract

Web-crawled corpora now claim that they have coverage of more than a thousand languages, but their subsets for endangered languages remain very small. Prior audits examine what is inside these corpora. We ask a question that complements this idea: what happens to the endangered-language web content that does exist? We curate and verify an inventory of 40 pages from active institutional websites (tribal language programs, community dictionaries, and learning portals) for 16 endangered North American languages from 7 families, and we attribute losses along the full pipeline: (S1) whether Common Crawl ever crawled the page, (S2) whether language identification (LID) recognizes the captured text, (S3) whether a standard quality-filter cascade can help retain it, and (S4) whether any of it reaches a released corpus (FineWeb-2, GlotCC). Most of the loss happens before quality filtering is even reached. These websites present to crawlers almost entirely in English, much of the actual language text is served by client-rendered dictionary applications that expose no text to crawlers, 6 of 16 languages have no label in the standard LID model, and none of the 36 inventory domains contributes a single document to any released per-language subset we examined. We release the inventory, the pipeline, and the per-stage attribution to support both crawl-side and community-side remedies.

1 Introduction

Most accounts of low-resource data scarcity focus on the corpus side. Quality audits show that web-mined subsets for low-resource languages are noisy or mislabeled (Kreutzer et al., 2022), and better mining recovers more (Dent et al., 2025). This paper examines the pipeline from the other end. Endangered languages with active revitalization movements have a real web presence, including language departments of sovereign nations, community dictionaries with audio, and lesson portals. If the released corpora for these languages are tiny, the loss must occur somewhere between that web presence and the corpus. This paper locates where.

We start from an independently curated, script-verified inventory (§3) of 40 pages across 16 endangered North American languages and trace each page through four loss stages (§4): Common Crawl capture (S1), language identification of the captured payload (S2), a FineWeb-2-style quality-filter cascade (S3), and presence in released corpora (S4). All stages are measured with public artifacts, namely the Common Crawl capture index (CDX), the underlying Web ARChive (WARC) records, GlotLID v3 (Kargaran et al., 2023), and the released FineWeb-2 (Penedo et al., 2025) and GlotCC (Kargaran et al., 2024) subsets. All code, the inventory with per-URL provenance, and the stage-level results are open source.[1]

2 Related Work

Corpus-side audits established that low-resource subsets of web corpora are often erroneous (Kreutzer et al., 2022). LID for low-resource languages has improved in coverage (Kargaran et al., 2023; Burchell et al., 2023) and in precision on closely related varieties (Fedorova et al., 2026), and rare-language extraction from Common Crawl has been reframed as targeted mining (Dent et al., 2025). Quality-filter cascades derive from C4 (Raffel et al., 2020) and Gopher (Rae et al., 2021). Our work complements this line of research by attributing where existing endangered-language content is lost before any mining begins. Our framing follows calls for community-centred data practice (Bird, 2020; Carroll et al., 2020).

3 The Inventory (S0)

Curation. We selected 16 endangered North American languages spanning 7 families and,

[1] https://github.com/Sharvin-coder/crawl-gap

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 220–223
October 29, 2026 ©2026 Association for Computational Linguistics

through documented web searches, collected 40 candidate URLs on 36 domains. The sites include official language programs of sovereign nations, community and university dictionaries, and learning portals. Every entry records its provenance, and no URL was invented or taken from any corpus. The inventory is a census of the verified pages we could locate, not a sample from a larger population, so we make no statistical generalization; the value of tracing it lies in attributing each individual loss.

Verification. A scripted pass fetched each page politely, extracted the server-visible text, and ran GlotLID v3. 37 of 40 pages are live. Three findings emerge at this stage. First, the document-level top-1 label of every live page is English. Second, the major dictionaries (FirstVoices, Cherokee Dictionary, itwêwina, the Blackfoot atlas) are client-rendered applications whose server-side HTML contains almost no text, so their content is invisible to any crawler. Much of what these sites publish is also not HTML at all: the pages we fetched link 299 distinct PDF files across 14 of 33 live domains, content a text-extraction pipeline does not see even when the linking page is crawled. Third, after following content-related internal links one hop deep and scoring the best page per site, only 11 of 34 language and site pairs expose any machine-detectable running text in the target language.

Accounting. One set of counts is used throughout. The curated inventory has 40 URLs on 36 domains, 37 of them live. The one-hop deepening step discovers 5 additional content pages not in the curated list, so the attribution pipeline traces 45 URLs in total. The funnel table (Table 1) covers only the 11 content pages with verified target-language text, which fall in 8 of the 16 languages.

4 Attribution Pipeline

For every traced URL we determine the first stage at which it is lost.

S1 (crawled?) We query the Common Crawl CDX index for exact-URL captures and host-level presence. Because a static page crawled once years ago would still enter FineWeb-2, a handful of recent crawls is not sufficient; we check 41 crawl releases spanning 2013 to 2026, including every release from 2023 through 2025, the newest 2026 releases, and one release per year back to 2013, which covers every crawl year FineWeb-2 draws on. We also check each host's live `robots.txt` policy for CCBot, the Common Crawl crawler.

Language	S1 never	S1 no-200	S2 no-label	S2 miss	S3 killed	S4 absent	S4 present
Arapaho (arp)	0	0	0	1	0	0	0
Blackfoot (bla)	0	0	0	1	0	0	0
Central Alaskan Yup'ik (esu)	0	0	0	1	0	0	0
Cherokee (chr)	0	0	0	2	0	0	0
Muscogee (mus)	0	0	0	0	1	0	0
Navajo (nav)	1	0	0	0	0	0	0
Ojibwe (ciw)	0	0	2	0	0	0	0
Plains Cree (crk)	0	0	0	2	0	0	0
Total	1	0	2	7	1	0	0

Table 1: First-failure attribution for content pages. Cells count pages (the best verified page per language and site); only the 8 languages with at least one page of verified target-language text appear, and the total is 11.

S2 (recognized?) For each URL with an HTTP-200 capture, we retrieve that WARC record by range request, extract its text, and run GlotLID v3. A page passes if the document-level top-1 label belongs to the target language, or if at least 10% of its text lines receive such a label with $p \geq 0.5$. Languages absent from the label space cannot pass this stage (6 of 16 languages: cic, ciw, lkt, mez, osa, tli).

S3 (retained?) We apply a re-implemented FineWeb-2-style cascade to the captured text, in a fixed order: a language-probability gate (document-level GlotLID probability at least 0.65), Gopher-style quality and repetition checks (minimum word count, alphabetic-word fraction, duplicate-line fraction, top character-bigram share, and bullet-line ratio) (Rae et al., 2021), and the C4 terminal-punctuation heuristic (Raffel et al., 2020). We record the first filter that fires; all thresholds are pinned in the released code.

S4 (released?) We scan the complete released per-language subsets of FineWeb-2 and GlotCC for documents from any inventory domain.

5 Results

Table 1 and Figure 1 give the funnel; per-stage detail follows.

S1. Of the 45 traced URLs, 8 have no capture in any of the 41 crawls (every index query completed), 1 were captured only with non-success HTTP statuses, and 36 have at least one successful capture; the three groups sum to 45. Only 1 of 36 hosts blocks CCBot in `robots.txt`, so the coverage gap comes almost entirely from discovery rather than from sites opting out. Never captured does not mean undiscoverable: these pages sit on live sites and are linked from pages we reached by ordinary browsing, so the gap reflects crawl

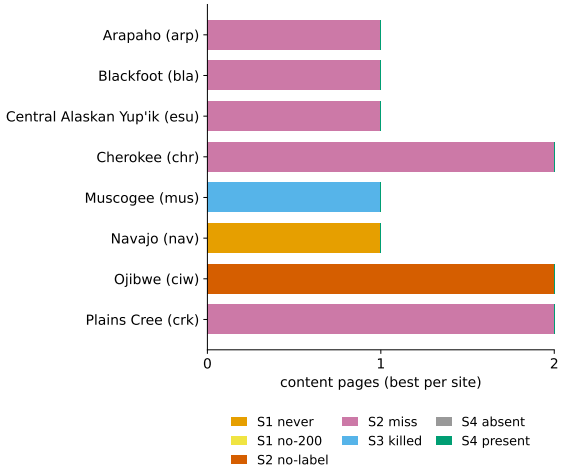


Figure 1: Stage attribution per language, for the same 11 content pages as Table 1.

scheduling and prioritization rather than technical unreachability. Crawl depth matters greatly here: 10 of the 36 captured URLs have their only successful captures outside the six most recent releases, so a shallow check against recent crawls alone would have roughly doubled the apparent never-crawled count.

S2. Of the captured pages, 1 passes, 23 are assigned to other labels, and 12 belong to languages with no label at all. The routing is not toward other minority languages: 23 of the 23 routed pages receive English as their top-1 label, because the target-language material is embedded in English framing. The languages missing from GlotLID's label space are exactly the languages that have no released subset in either corpus.

S3. Applied to all captured pages, the cascade keeps 20 and removes 16, most often at the language-probability gate. The single page that passed S2, a bilingual lesson page, is among those removed: its English framing pulls the document-level language probability below the gate. Because revitalization materials are typically bilingual, this behavior affects them systematically.

S4. The released subsets are small. Across the 10 of 16 languages that have a per-language subset at all, the subsets scanned here range from 1 to 239 documents (976 in total). The number of documents from inventory domains in either released subset is 0. This establishes absence from the two corpora we can scan document by document. As a breadth check on label coverage, none of the 16 languages has any per-language subset in mC4, OSCAR-23.01, or CulturaX, and MADLAD-400 carries labels for 4 of the 16 (Cherokee, Navajo, Ojibwe, and Cree) and none for the other 12. The institutional web presence of these 16 languages contributes no documents to the corpora examined here.

6 Discussion

These losses are structural rather than incidental. Institutional pages are written in English for good reasons, since their audience includes learners and citizens who do not yet speak the language. The most valuable language content sits behind JavaScript applications and PDFs. LID label spaces were built from languages already observable in corpora, and this circularity keeps unlabeled languages out. Remedies need to match the stage where the loss happens. On the crawl side, rendering support or capture agreements with community dictionaries would expose text that currently stays hidden. Line- or paragraph-level LID ahead of document-level gating would let bilingual pages contribute their target-language lines instead of failing as a whole. Label spaces can grow from inventories instead of corpora, and verified community domains can be exempted from filters that were tuned for general web text. Above all, communities should keep authority over whether and how their language data enters corpora (Carroll et al., 2020; Bird, 2020).

Limitations

Our inventory covers one region (North America), 16 languages, and 45 URLs; it is small because verifiable target-language pages are genuinely scarce, and we trace the population we found rather than sampling from it. A matched control set of institutional pages in higher-resource languages, and expansion to more languages and domains in collaboration with community partners, are natural next steps. The attribution thresholds (the S2 pass criterion and the S3 cascade) are pinned but necessarily approximate the released pipelines. PDF and audio content, which makes up a large share of what language programs publish, is out of scope for the text funnel, so our estimate of the crawlable text is conservative; §3 quantifies the PDF surface. Some inventory domains may be represented in corpora under other languages' subsets (typically English), which we do not count as presence for

the target language.

Ethics Statement

We fetch only publicly served pages, at most once per page per run, with an identifying research user agent and per-host delays. The repository releases URLs, metadata, and derived statistics, and never scraped text, consistent with the CARE principles (Carroll et al., 2020). The released artifact documents a correction-and-removal process, and we commit to honoring removal or correction requests from the nations or from site stewards. Corpus presence of community content is reported in aggregate. We do not redistribute any community language material.

References

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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)*, Toronto, Canada. Association for Computational Linguistics.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Rasul Dent, Pedro Ortiz Suarez, Thibault Clérice, and Benoît Sagot. 2025. Identifying rare languages in Common Crawl data is a needles-in-a-haystack problem. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1460–1473, Suzhou, China. Association for 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 (VarDial)*, pages 275–292, Rabat, Morocco. Association for Computational Linguistics.

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*, Singapore. Association for Computational Linguistics.

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 (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, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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.

You Can Tell Who's Asking: What the Web's Questions Are Made Of, and Where They Come From

Calvin Zhou[1,2]* **Vincent McCloskey**[2] **Krishna Srinivasan**[2]
[1]University of Pennsylvania [2]Bodhium Labs
{calvin, vincent, krishna}@bodhiumlabs.com

Abstract

Questions scraped from the web are used across academia and industry as a proxy for what people want to know. Across QA training data, retrieval benchmarks, and content strategy, questions on a page are assumed to reflect human intent. We test this assumption at scale by extracting 13.4B question occurrences across 110 FineWeb snapshots (2013-2025), and report three findings. First, you can tell who is asking: provenance (the host/page of questions) leaves a signal in question form, and a logistic model can separate genuine user questions from templated/manufactured ones at AUC 0.725 via length and surrounding context rather than question type, though only 0.554 against commerce FAQ writing. Second, question frequency does not measure demand: the most-frequent questions are boilerplate/templated (over 70% of the top thousand), so occurrence counts measure how often a string was published and not how often it was asked. Third, over twelve years the genuine share of occurrences fell by 79% (42-56% after controlling for crawl composition), with question length and context decreasing. We present the first diachronic, occurrence-level measurement of web question provenance, and find the crawlable web's questions have shifted from being asked by humans toward manufactured for machines to read.

1 Introduction

A question mark on a page was never a guarantee that someone genuinely wanted an answer, but it was a reasonable bet. For years, that assumption has weakened as search engines rewarded pages for including questions. FAQ blocks were attached to the bottom of pages to rank higher, with a rise evident in *FAQPage* markup adoption (Brinkmann et al., 2023; HTTP Archive, 2024). That reward has diminished since then (HTTP Archive, 2024), but in the Generative Engine Optimization (GEO) era, content more than ever is being written for machines (Aggarwal et al., 2024), with machine-generated text making up a growing share of the crawlable web since late 2022 (He et al., 2026; Dolezal et al., 2026). Yet, extracted web questions are still seen as what people want to know in QA training data (Huber et al., 2022; Dinzinger et al., 2025), retrieval benchmarks (Nguyen et al., 2016; Kwiatkowski et al., 2019), and content strategy.

What a question dataset can represent is dependent on its source (Gebru et al., 2021; Dodge et al., 2021). Datasets of genuine questions, such as query logs and community-QA archives, are small and kept in-house, with the AOL log with true query frequencies even being recalled (Pass et al., 2006). Web-scale datasets, whether FAQ-markup mined or autocomplete-harvested, include both the web's genuine and manufactured sources (App. A). Mining full page text at least provides occurrence counts, but a count of appearances is not a count of asks. The scarce but valuable signal is *popularity*, which turns a list of questions into a distribution.

We measure the web's question layer from 2013-2025 across FineWeb (Penedo et al., 2024), with *question occurrence* as our unit: a single question-bearing sentence on a crawled page. We ask whether a question's provenance is determinable from its form (§4), whether how often it appears relates to demand (§5), and how the proportions of user-generated content (UGC) and manufactured questions changed across the SEO, COVID, and AI eras (§§6-7).

2 Related work

Web Data Commons, the closest prior work, extracts schema.org markup from one Common Crawl snapshot per year and shows FAQPage/Question adoption increasing rapidly (Brinkmann et al., 2023; Meusel et al., 2015),

*Work done during an internship at Bodhium Labs.

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 224–237
October 29, 2026 ©2026 Association for Computational Linguistics

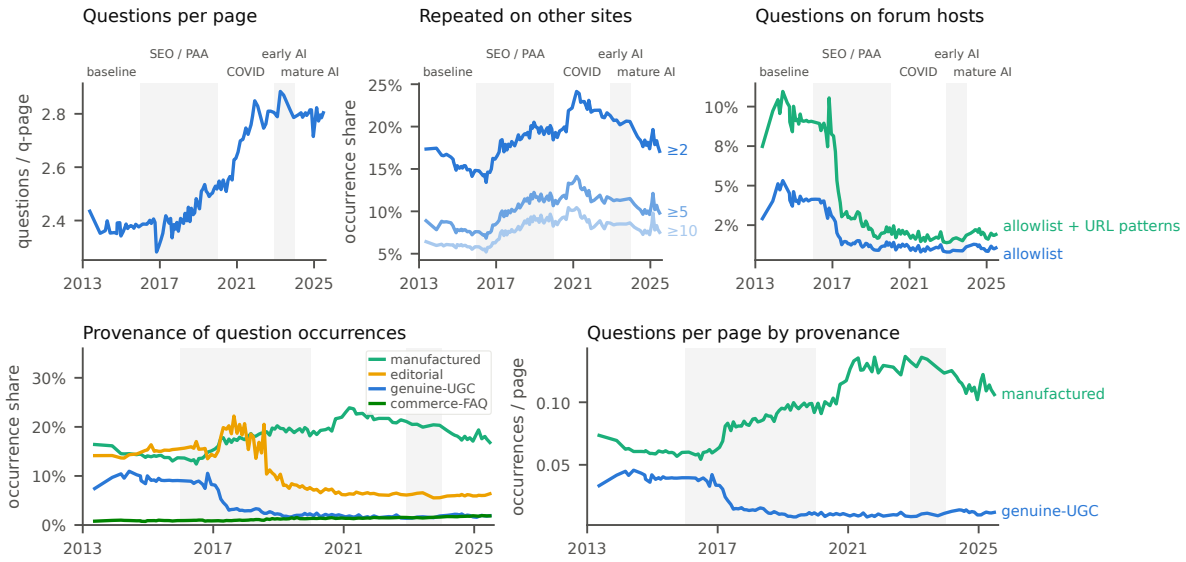


Figure 1: Twelve-year changes over 110 snapshots (§6). Top: question intensity, cross-host duplication ($D \geq 2/5/10$), and genuine-forum share (allowlist and allowlist + URL patterns). Bottom: the provenance classes (unknown omitted), first as a share of occurrences, then as questions per crawled page (does not depend on the size of unknown). Shaded bands are the five eras, named along the top row. Trends use the Hamed-Rao correction.

driven first by SEO then AI-search (HTTP Archive, 2024). While they count markup classes by domain, we count linguistic questions, which are the schema-less majority, and label each with provenance and form. CCQA (Huber et al., 2022) mines schema.org question markup, WebFAQ (Dinzinger et al., 2025) only FAQPage (2.0 is re-crawled (Dinzinger et al., 2026)), both of which treat the result as natural asking; GooAQ (Khashabi et al., 2021) harvests autocomplete; PAQ (Lewis et al., 2021) is machine-generated. No prior work analyzes provenance or temporal changes at the question level.

Prior work audits web document LLM-content prevalence (He et al., 2026), the machine-generated text share over time (Dolezal et al., 2026; Sun et al., 2025; Thompson et al., 2024), and mass-produced SEO spam in search (Bevendorff et al., 2024), but none reaches the question layer. We also add a signal that none use: verbatim cross-host duplication (text that appears word-for-word on multiple different websites), which decreases in our study. Since others report machine text increasing, this is detector-independent evidence of a change in how questions are manufactured (§6). Change in question form was last measured query-side through 2010 (Pang and Kumar, 2011), on queries typed into a search engine rather than questions written on pages, and without provenance labels; query-side work has also compared question and keyword phrasings (White et al., 2015). We measure it page-side, at the level of the individual occurrence, and label where each one came from. Corpus-composition auditing at scale is the closest type of work (Elazar et al., 2024), as is the human audit of web-crawled corpus quality in Kreutzer et al. (2022); neither reaches the question layer.

3 Data and methods

Corpus. We analyze questions from 110 English, quality-filtered FineWeb dumps, CC-MAIN-2013-20 through CC-MAIN-2025-26, one per Common Crawl crawl snapshot.[1] A *question occurrence* is a sentence with a question mark of 5-100 words that starts with a question word or contains an interrogative keyword, extracted from page text and labeled with URL, host, and snapshot. We normalize before counting (NFKC, casefold, whitespace collapse, edge-punctuation strip), resulting in 13.4B occurrences with a mean length of 12.8 tokens. Our extractor was fixed and did not adapt to differences in era (see Limitations).

Design. We use two units of analysis throughout this work: every temporal curve and provenance

[1] Hugging Face dataset `HuggingFaceFW/fineweb`, revision `9bb295ddab0e05d785b879661af7260fed5140fc` (tag `v1.4.0`, 2025-07-11) (Penedo et al., 2025). Penedo et al. (2024) document the 96-dump release covering 2013 to April 2024; v1.4.0 extends it through CC-MAIN-2025-26 (June 2025) and contains the 110 dumps we process.

share is computed over *occurrences* (one question-bearing sentence on one crawled page), while the frequency analysis in §5 alone works on *distinct questions* (one normalized string within a snapshot). All statistics are first computed within each snapshot then compared to others after occurrence-weighting. We never sum counts across snapshots, and subsequent percentages are relative change. Trends use a two-sided Mann-Kendall (MK) test (Mann, 1945; Kendall, 1975), and comparisons use five prior-frozen eras: baseline (2013-15), SEO (2016-19), COVID (2020 through CC-MAIN-2022-40), early-AI (CC-MAIN-2022-49, the ChatGPT release week, through 2023), and mature-AI (2024-25), with bootstrap CIs over per-snapshot shares (the widest era-mean half-width below is 0.8 points) and a $\pm$1-year boundary check. Consecutive snapshots are autocorrelated, so every trend below is reported under the Hamed-Rao variance correction (Hamed and Rao, 1998), which cuts the effective series length to $n^*\approx$9-13; a series we describe as non-monotonic gets no whole-series trend test, since MK tests monotonic trend only. We also correct for testing many curves (App. F). Robustness checks for crawl-size variation are in Limitations.

Metrics. *Intensity* is occurrences per question-bearing page, which is the quantity plotted throughout. The share of pages with any question increases over the series, so this and the per-crawled-page rate diverge, and we report both. *Cross-host duplication* is the share of occurrences whose normalized text appears on at least D registrable domains within a snapshot, which is a conservative lower bound at $D\geq2$ (we do no fuzzy matching; same shape at $D\geq5$ and $D\geq10$; see Limitations on snapshot size and detection). *Genuine-forum share* classifies hosts with a precision-prioritized UGC classifier, an allowlist of person-to-person hosts and URL patterns like `/forum/`, and we report separately the allowlist and allowlist + URL patterns union. *Provenance* assigns each occurrence to one of five classes: genuine-UGC, manufactured-template (cross-host duplicated), commerce-FAQ, editorial, or unknown (App. B). *Form* is a deterministic regex and part-of-speech (POS) classifier (App. C) over type, person, and scaffolding (an attached context sentence), validated against human labels (App. D). Annotators agree on provenance at κ=0.52 but on authenticity (whether a question was genuinely asked) at only κ=0.23, which is why we focus on provenance and not authorship. We run no machine-generated-text detector since our unit is far too short for reliable detection and such detectors are fragile even on long text (Sadasivan et al., 2025), so duplication, provenance, and form stay detector-independent. However, this leaves a large unknown residual (61%-73% over the series; see Limitations).

Sampling. The occurrence-weighted curves use the full population. Form and context being per-question properties, are estimated on a per-question sample: 20k questions per snapshot (2.2M total), drawn via a deterministic hash of the question text from files spaced across each dump. At this size a per-snapshot share carries 95% CI half-width under 0.7 points, so the aggregated sample keeps the provenance strata large enough (28k commerce-FAQ, 92k genuine-UGC) for the form-provenance analysis in §4.

4 Form predicts provenance

On a 2.2M sample with both form and provenance labels (Table 1), genuine questions are longer than manufactured ones (12.9 vs. 8.5 tokens, Cohen's d=0.71), about twice as likely to be scaffolded (7.3% vs. 3.9%), and more often to be first-person (29.7% vs. 21.4%). A logistic regression on only form features separates the two at AUC 0.725, and length is the dominant term (odds ratio 2.09 per SD). This is a population-level separation, since at the level of a single question, the model is weak, and we make no claim about individual items. It is also narrower than the pooled figure suggests. Against the commerce-FAQ class, the only class defined without duplication, AUC falls to 0.554 (App. B), so what form separates is genuine questions from template farms, not from commercial-FAQ writing, where only provenance gives them away. As an external check, WebFAQ, mined from *FAQPage* markup, matches our manufactured class on every form measure (App. C).

We pre-specified five hypotheses before analysis (App. E). The one that failed was question *type*, such as wh- and polar, which has almost no signal (Cramér's V=0.03); manufactured questions are stripped of length and context, and of first-person framing only weakly (V=0.04). This holds across the eras, as genuine questions stay 12.8-13.4 tokens in all, the only class whose form does not decline (App. C).

	genuine	manuf.	effect
length (tokens)	12.9	8.5	d=0.71
scaffolded	7.3%	3.9%	ϕ=0.07
first-person	29.7%	21.4%	ϕ=0.08

Table 1: Form by provenance; 2.2M sample (n=511k: 92k genuine, 420k manufactured). *Scaffolded*: the question has an attached context sentence. Manufactured = template ∪ commerce-FAQ. Effects are genuine vs. manufactured, where length does most of the work; App. E reports Cramér's V for person and type over all five classes.

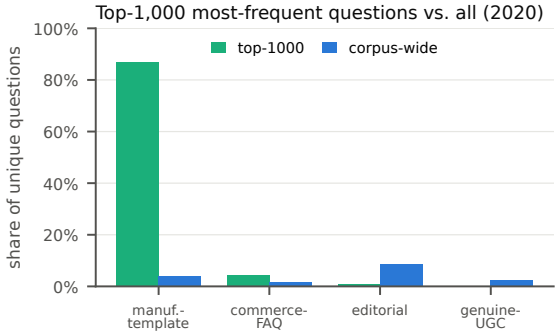


Figure 2: Provenance of the top-1,000 most-frequent questions vs. all distinct questions (2020; *unknown* omitted). Of the top-1,000, 86.8% are cross-host duplicated strings, against 4.1% of distinct questions corpus-wide.

5 Frequency is not a reliable proxy for demand

A natural way to read demand from the web is to count, but we find this to be flawed. All counts here are within-snapshot occurrence counts of normalized text, on three snapshots: 2015, 2020, 2024 (one in the middle of each era).

We rank each snapshot's distinct questions by count. The top-1,000 are 71.7%, 86.8%, and 93.2% manufactured-template in 2015/2020/2024, against 3-4% of distinct questions corpus-wide. The genuine-UGC share of that is 6.3% in 2015 and 0.3% in both 2020 and 2024. The most-verbatim-repeated strings are boilerplate: "what are you waiting for" (72,331 occurrences in 2020), "how can we help you". Furthermore, genuine demand never concentrates in the distribution, with 97% of distinct questions occurring once or twice and the top-1,000 are under 3% of all occurrences. Therefore, occurrence frequency measures publishing intensity. We make no claim in this work about what real demand is.

6 Who's asking, over time

Figure 1 summarizes the twelve-year changes.

Intensity rises from an era mean of 2.37 in 2013-15 to a 2.83 peak in 2022-23, settling at 2.79 by 2024-25 (+17.7% over twelve years), concentrated in the SEO and COVID eras when FAQ adoption accelerated as previously mentioned. The share of pages with any question (18.4% to 22.5%, 2013-2025) moves independently of FineWeb's per-snapshot retention (Spearman $\rho = -0.08$). Since more pages have questions and each has more of them, questions per crawled page increase +46.6%.

Cross-host duplication is non-monotonic: 15.7% of occurrences at baseline, up to 21.6% at a peak in CC-MAIN-2021-10, then down to 18.3% through the AI eras (same shape at $D \geq 5/10$). We therefore do not report a whole-series trend for this curve. Duplication is also a detection event, and snapshot size varies 3.6× over the series, so we hold size fixed: each mature-AI snapshot duplicates less than the 2020-21 snapshot nearest it in size, 16 times out of 16 (mean −3.5 points). The peak precedes the ChatGPT release by around twenty months. If questions were generated by simply copying templates, duplication would keep increasing. Instead, AI-era text appears novel, so verbatim duplication would undercount manufactured questions after 2022. Detector-based studies find machine content rising over the same window (He et al., 2026; Dolezal et al., 2026).

Genuine-forum share collapses on both definitions: the allowlist from 4.2% to 1.1% (−74.3%) and the broader allowlist + URL pattern union from 9.4% to 2.0% (−79.0%). On a fixed panel of frequent hosts, the declines shrink to −42%/−56% (see Limitations). Its recall on human-labelled UGC is lowest in the baseline era (App. D), which would understate the decline rather than inflate it. Most of the fall precedes the 2023-24 crawler-blocking wave (Longpre et al., 2024) and holds on pre-2023 snapshots, so it is not confounded by forums leaving the crawlable web (see Limitations).

Provenance. From 2013 to 2025, genuine-UGC falls from 9.3% to 1.9% (−79.2%), manufactured-template rises +25.9% (same peak as the duplication curve), and commerce-FAQ doubles from a small base (+102%). These are shares of all occurrences, so a growing unknown class would shrink them on its own. We therefore also measure each class per page: genuine-UGC falls from 0.222

to 0.054 occurrences per question-bearing page (-75.5%), and from 0.040 to 0.012 per crawled page (-69.5%), while manufactured-template rises $+84.6\%$ and commerce-FAQ $+196\%$ per crawled page. Unknown accounts for at most ten of the seventy-nine points, so renormalization is not the cause of the decline. Crawl size is a confound, and adjusting for it, the manufactured-template *share* rise does not survive (-2.0%, p=0.49) while the per-page rate does ($+13.3\%$, p=0.006). We therefore claim the growth of the manufactured layer as a rate and not as a share (see Limitations). This is corroborated by a simultaneous rise in FAQ-path URLs (`/faq`, `questions-and-answers`) from 0.82% to 1.42% (z=+6.6), steepest from 2017-19, the same window of rapid *FAQPage* adoption. However, this counters the idea of "question farms", as the top-100 hosts' share *falls* from 11.7% to 2.8%.

Form. Mean length falls from 13.41 to 11.85 tokens (-11.6%) and the scaffolded share from 9.95% to 6.12% (-38.5%). What diminishes is the human around the question, such as a background sentence or first-person framing, which is the same pattern we see in §4. The remaining genuine questions are verbose while the growth is in stripped templated strings, matching studies of online Q&A community volume collapsing (del Rio-Chanona et al., 2024; Burtch et al., 2024); on Stack Overflow, the questions that still get asked are longer and harder to answer (Helic and Santos, 2026).

7 Context-independence

Most of the questions we observe are not self-contained; under a deterministic POS-based rubric (self-contained / conversational-fragment / context-dependent), about half of our sample are self-contained, 38% are conversational fragments, and 12% carry unresolved references. A human check agrees with the rubric on 63% of items, and the needs-context flag has precision 0.75 (App. H), so we read this section as exploratory and its rates as directional. Over time, the context-dependent share decreases from 14.0% to 8.6%, which holds under the Hamed-Rao correction; the matching rise in the self-contained share does not (p=0.20), so we do not make a claim about it.

Genuine-UGC is the most context-dependent class (19%); we interpret this as real questions being written for people to read, so they depend on surrounding context. Commerce-FAQ is the most self-contained (59%), because an FAQ page's purpose is to present and answer a standalone question. Filtering a web corpus for clean, standalone questions therefore preferentially keeps the manufactured questions.

8 Discussion

A growing share of the web's readable questions is there to be matched by search engines and now generative answer engines. Systems that interpret web question frequency as demand at face value are sampling the manufactured layer of short, templated, impersonal, standalone strings. When retrieval for such systems draws on this layer, quality has been shown to degrade in simulation (Yu et al., 2026).

The manufactured layer is full of fluent and well-formed text on the surface. Telling the genuine from manufactured requires looking at provenance, i.e. where a question occurs, how common it is, what form it takes, and its context. Frequency does not measure demand on its own (§5).

Today's manufactured web was built by the SEO economy well before 2022. After that we see the signal weaken: verbatim cross-host duplication, which caught manufactured questions for a decade, catches less of them in the AI eras, and its peak precedes the ChatGPT release by around twenty months. We run no detector, so we cannot attribute that to model-written text. Either way, duplication no longer finds the manufactured layer, and future measurement will need a different signal.

Limitations

Corpus scope. Our curves are measured on the FineWeb-filtered English web. A re-extraction from raw Common Crawl WET text (three dumps spanning the range: CC-MAIN-2015-40, -2020-34, and -2024-51) reproduces the twelve-year directions web-wide (intensity up, forum share down from a 12.7% raw base in 2015), but the raw and filtered curves diverge after 2020. We therefore make the twelve-year claims web-wide and confine the post-2020 contrasts (early-AI vs. mature-AI) to the FineWeb slice. Our analysis is restricted to English; trends may differ across languages.

Filter-induced exclusion. Our design cannot separate a decline in genuine UGC on the filtered web from a quality filter that increasingly excludes it. The raw 2015 forum base is 12.7% against 9.4% filtered, and quality filters are documented

to select against informal registers and community-authored text (Dodge et al., 2021; Kreutzer et al., 2022). Part of what we measure as genuine UGC leaving the corpus may be the filter declining to keep it. This matters beyond our estimates, because the excluded material skews toward the same non-standard registers the field has been trying to stop under-representing. The raw re-extraction that bounds this rests on three dumps, which is enough to check direction and not enough to correct magnitude.

Measurement. Our genuine-UGC detection values precision over recall, so the genuine shares are lower bounds and we claim trends only and genuine text on unrecognized hosts is left to the *unknown* class. Shares are computed over all occurrences, so a growing unknown residual shrinks every classified share (Aitchison, 1982), and part of a measured decline could be questions moving into unknown. Measured per page instead, genuine-UGC still falls -75.5% per question-bearing page and -69.5% per crawled page against the -79.2% share, so the residual accounts for at most ten of those points. Of 49 human-labelled unknown items only 4 were genuinely UGC; reassigning that 8% of the residual to genuine gives a -45.2% decline that still holds under the correction (p=0.01), and the decline would only disappear if a quarter of the residual were genuine, which is three times what the sample supports. Unknown's form also de-scaffolds over time like the manufactured and editorial classes, while genuine UGC stays flat. The unknown class remains the weakest part of the scheme: it holds 61%-73% of occurrences at precision 0.43, so we treat every classified share as a lower bound rather than a partition of the web's questions. Because we run no short-text machine-generated-text detector (§3), the manufactured class inherits verbatim duplication's post-2022 undercounting, and we claim only that our AI-era findings are consistent with a shift to AI-generated text and not that any question was model-written. The fixed extractor cannot adapt to era-specific styles, so recall drift could in principle mimic a trend; but the forms it undercaptures (elliptical and noisy informal questions) skew genuine, so its misses leave our genuine-share estimates conservative. Form labels are heuristic and evaluated against human labels (App. D).

Crawl artifacts. The host-level measurements could move with what Common Crawl fetches rather than with the web. Which hosts it fetches shifts over the years (Thompson, 2024); on a fixed panel of the 79,826 hosts present in ≥90% of snapshots the forum declines are still the same direction but shrink (-42%/-56% vs. -74%/-79% crawl-wide) and top-100 concentration falls -57%, so the direction is a genuine web trend. The decline is smaller on the panel, and that gap, about 40% of the full-crawl figure, comes from changes in which hosts Common Crawl fetches. Major UGC platforms have also restricted open crawlers since 2023 (Longpre et al., 2024), so late-era genuine levels reflect only the crawlable web, but recomputing on pre-2023 snapshots shows withdrawal does not drive the decline and is if anything steeper. Consecutive snapshots are autocorrelated since under the Hamed-Rao correction the effective series length falls to n^*≈9-13, yet intensity, both forum signals, and shallow duplication ($D\geq2$) stay significant ($p < 0.02$), while $D\geq5$ is marginal and $D\geq10$ non-significant. Thus, we claim duplication only at the shallow threshold.

Crawl size. Snapshot size varies 3.6× over the series (57.6M to 206M occurrences), and duplication is a detection event, so larger snapshots find more of it: the first differences of the duplication and size series correlate at ρ=+0.66, against $|\rho|\leq0.33$ for every other curve. Regressing each curve on log size and re-levelling at the mean, the twelve-year contrasts shrink and two reverse. The apparent net increase in duplication does not survive (from $+16.6\%$ to -4.7%, p=0.81), and neither does the manufactured-template share's (from $+25.9\%$ to -2.0%, p=0.49); we make neither claim. What survives is intensity ($+6.9\%$), commerce-FAQ ($+46.4\%$), the manufactured-template rate per crawled page ($+13.3\%$, p=0.006), top-100 concentration (-39%), the form drift, and the genuine decline at -45.0%, which is where the fixed host panel also lands (-42%/-56%). Two independent controls converging on roughly -45% is our best estimate of the genuine decline; the crawl-wide -79% is an upper bound on it. The size-adjusted genuine series does not clear the Hamed-Rao threshold on its own (p=0.13). With n^*≈9 and two corrections applied in sequence, we are at the limit of what 110 snapshots can support.

Ethics and artifacts

The corpus is derived from Common Crawl via FineWeb (ODC-By). We analyze public web text in aggregate and release no personal data. The valida-

tion sample is constructed to be PII-clean, with candidates containing a person name, email, or phone number being dropped at selection and replaced from the same stratum, and was screened for sensitive topics by a keyword filter plus a manual read of every item. Host and URL are withheld from every released artifact, so no released string is linkable to a page or author. At `https://github.com/bodhiumlabs/tell-whos-asking` we release the validation sample and its guidelines, the UGC host classifier (App. G), the commerce and editorial host lists and hostname patterns, the interpretable form and context classifiers, the LLM-judge script (prompt, model id, and decoding settings), per-snapshot denominators, the fixed host panel, per-figure curve data, and the deterministic rule that selects the 2.2M sample (the fingerprint-modulus predicate over question text, the modulus, and the per-dump file spacing), which rebuilds the same sample from public FineWeb without releasing document identifiers. The per-occurrence text stays unreleased. The validation sample contains no content from hosts that signalled crawler opt-out at the time of release (Longpre et al., 2024).

Funding and competing interests. All authors are affiliated with Bodhium Labs, which develops commercial products in answer- and generative-engine optimization. The work was funded by Bodhium Labs, which also provided the compute.

Dual use. A released classifier that separates genuine from manufactured questions can also be used to evade it. The changes §4 implies are cheap with an LLM: lengthen the question, attach a context sentence, write in the first person. If content farms make them, this measurement stops working, so we expect it to date quickly. We release anyway because the same detail lets dataset builders audit what they have collected. Read the other way, our decline covers the crawlable, quality-filtered layer only, and taking it as a verdict on community-generated content would compound the exclusion described above.

Acknowledgments

We thank Ruben Ramirez Salas and Alex Miller for their comments on earlier drafts, and the four anonymous reviewers for the depth of their reviews.

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024)*.

John Aitchison. 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B*, 44(2):139–177.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. 2024. Is Google getting worse? a longitudinal investigation of SEO spam in search engines. In *Advances in Information Retrieval (ECIR 2024)*.

Alexander Brinkmann, Anna Primpeli, and Christian Bizer. 2023. The web data commons schema.org data set series. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*.

Gordon Burtch, Dokyun Lee, and Zhichen Chen. 2024. The consequences of generative AI for online knowledge communities. *Scientific Reports*, 14:10413.

Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul Bennett, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, Wenqi Jiang, Zheng Liu, Mingqin Li, and 12 others. 2024. MS MARCO web search: A large-scale information-rich web dataset with millions of real click labels. In *Companion Proceedings of the ACM Web Conference (WWW)*.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 18 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020)*, pages 2983–2989.

R. Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. 2024. Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus*, 3(9).

Michael Dinzinger, Laura Caspari, Kanishka Ghosh Dastidar, Jelena Mitrović, and Michael Granitzer. 2025. WebFAQ: A multilingual collection of natural Q&A datasets for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Michael Dinzinger, Laura Caspari, Ali Salman, Irvin Topi, Jelena Mitrović, and Michael Granitzer. 2026. WebFAQ 2.0: A multilingual QA dataset with mined hard negatives for dense retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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 (EMNLP 2021)*.

Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. The impact of AI-generated text on the internet. *Preprint*, arXiv:2604.26965.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Khaled H. Hamed and A. Ramachandra Rao. 1998. A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204(1–4):182–196.

Sichang Steven He, Calvin Ardi, Ramesh Govindan, and Harsha V. Madhyastha. 2026. DeGenTWeb: A first look at LLM-dominant websites. *Preprint*, arXiv:2605.00087.

Denis Helic and Tiago Santos. 2026. Stack Overflow is not dead yet: Crowd answers still matter. *Journal of Systems and Software*, 241:112988.

HTTP Archive. 2024. Web almanac 2024: Structured data. `https://almanac.httparchive.org/en/2024/structured-data`.

Patrick Huber, Armen Aghajanyan, Barlas Oğuz, Dmytro Okhonko, Wen-tau Yih, Sonal Gupta, and Xilun Chen. 2022. CCQA: A new web-scale question answering dataset for model pre-training. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First Quora dataset release: Question pairs. Quora Engineering Blog.

Maurice G. Kendall. 1975. *Rank Correlation Methods*, 4th edition. Charles Griffin, London.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9.

Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Anis, An Dinh, Caroline Chitongo, Da Yin, and 30 others. 2024. Consent in crisis: The rapid decline of the AI data commons. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*.

Henry B. Mann. 1945. Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

Robert Meusel, Christian Bizer, and Heiko Paulheim. 2015. A web-scale study of the adoption and evolution of the schema.org vocabulary over time. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics (WIMS 2015)*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation (CoCo@NeurIPS 2016)*.

Bo Pang and Ravi Kumar. 2011. Search in the lost sense of "query": Question formulation in web search queries and its temporal changes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011), Short Papers*, pages 135–140.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems (InfoScale)*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. FineWeb (v1.4.0). Hugging Face dataset `HuggingFaceFW/fineweb`, revision `9bb295ddab0e05d785b879661af7260fed5140fc`.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*.

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are we in the AI-generated text world already? quantifying and monitoring AIGT on social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775.

Henry S. Thompson. 2024. Improved methodology for longitudinal web analytics using Common Crawl. In *Proceedings of the 16th ACM Web Science Conference (WebSci 2024)*.

Ryen W. White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. queries in informational search tasks. In *Proceedings of the 24th International Conference on World Wide Web (WWW 2015 Companion)*.

Yahoo! Research. 2007. Yahoo! answers comprehensive questions and answers, version 1.0 (webscope l6). Yahoo! Webscope Program.

Hongyeon Yu, Dongchan Kim, and Young-Bum Kim. 2026. Retrieval collapses when AI pollutes the web. In *Proceedings of The Web Conference 2026 (WWW '26)*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1m ChatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*.

A Question datasets

Table 2 places this work among representative question datasets by provenance family. Each family carries a distinct limitation, i.e. why its questions are not a clean record of human demand:

Query logs record real asking, but as *mediated clicks*; the true frequencies that would make them a distribution were withdrawn (AOL) or stripped (ORCAS).

Community-QA is human-authored but gated or deduplicated, and its vote signal ranks only within a single forum.

Markup mining takes site-published `schema.org` QA annotation: WebFAQ restricts to `FAQPage`, while CCQA's `Question` mixes those with community asking (Stack Exchange is its largest source), unlabelled.

Autocomplete/PAA harvests search-*suggested* completions, not questions anyone wrote directly.

Synthetic generation has no human denotation at all and is effectively unbounded.

Chat logs capture people *prompting an assistant*, not asking on the open web.

Full-text mining (this work) reads question sentences off pages at web scale, so it reaches the whole distribution, but its composition must be measured.

B Provenance classes

Each question occurrence is labeled by an *ordered cascade* over its host and URL where the first matching rule wins. In order, each test applying only to occurrences the previous ones did not take:

Dataset	Yr	Family	Size	Freq	Access
MS MARCO WS (Chen et al., 2024)	'24	query log	10^7	dist.	open
ORCAS (Craswell et al., 2020)	'20	query log	10^7	no	open
AOL log (Pass et al., 2006)	'06	query log	10^7	yes	withdrawn
Yahoo! L6 (Yahoo! Research, 2007)	'07	CQA	10^6	no	gated
Quora (Iyer et al., 2017)	'17	CQA	10^5	no	open
GooAQ (Khashabi et al., 2021)	'21	autocompl.	10^6	no	open
CCQA (Huber et al., 2022)	'22	QA markup	10^8	no	code
WebFAQ (Dinzinger et al., 2025)	'25	FAQ markup	10^8	no	open
PAQ (Lewis et al., 2021)	'21	synthetic	10^8	no	open
Persona Hub (Ge et al., 2024)	'24	synthetic	10^9	no	partial
WildChat (Zhao et al., 2024)	'24	chat log	10^6	no	terms
LMSYS-1M (Zheng et al., 2024)	'24	chat log	10^6	no	terms
This work	'25	full-text	10^{10}	counts*	aggr. only

Table 2: Question datasets by provenance family (compressed; representative members). **Size**: order of magnitude of questions/queries. **Freq**: whether the dataset carries a frequency/popularity signal; *yes* = per-query counts released, *dist.* = the sample preserves the query distribution but not per-query counts, *no* = a deduplicated list. *We report within-snapshot occurrence counts, which is not demand. *aggr. only*: we release per-figure curve data and the validation sample and not the per-occurrence text.

1. UGC host set, or person-to-person URL pattern? → `genuine-UGC`
2. normalized text on ≥ 2 registrable domains? → `manufactured-template`
3. host is commerce or support/FAQ? → `commerce-faq`
4. host is news, blog or media? → `editorial`
5. otherwise → `unknown`

The classes are therefore mutually exclusive by construction, and `unknown` is reached only by failing all four tests. Duplication is computed on the question text normalized by NFKC, casefold, whitespace collapse, and edge-punctuation stripping (as in §3). The rules in full:

`genuine-UGC` Host is in a user-generated-content set: the ∼1M most frequent hosts (by occurrence) run through a forum/UGC host classifier, *or* the URL matches a person-to-person pattern (`/forum`, `/thread(s)`, `/topic`, `/comments`, `/discussion`, `/board`, `viewtopic.php`, `showthread.php`, or a `forum.`/`ask.`/`community.`/`discuss.` subdomain). We tested this first, and it works even for duplicated text. This is a lower bound, since only recognized UGC hosts/patterns are caught, so genuine questions on unrecognized hosts fall through.

`manufactured-template` Among non-UGC pages, the normalized question occurs on ≥ 2 distinct registrable domains: cross-host duplication (canonical FAQ/People-Also-Ask/boilerplate republished across sites). This is defined by *duplication* and not by wording or length.

`commerce-faq` Non-UGC, non-duplicated; host is a commerce or support/FAQ site, identified first from a curated list of hosts labelled by category, and for hosts not on the list from hostname patterns (`shop|store|cart|buy|deals|coupon|pricing|ecommerce`; `support|help|faq|docs|kb|answers|knowledgebase`).

`editorial` Host is a news/blog/magazine/media outlet (`news|blog|magazine|times|post|herald|journal|media|wire`).

`unknown` No rule matched: not recognized UGC, not duplicated, and the host in no category. This is the residual, unclassifiable layer, whose share increases from 61% to 73% over the series; whether that shrinking base relabels the classified trends is in Limitations.

These hostname patterns match substrings, so some of them collide: `post|times|wire` also catches *poster* and *wireless*, and `answers|docs|kb` also catches community Q&A and technical documentation. We re-ran the host-type labeling with those alternatives dropped, on one snapshot per year. The editorial trend is unaffected (-63.1% against -64.6% pruned), while the FAQ trend roughly doubles ($+54.9\%$ against $+115.0\%$), so dropping the collisions strengthens the FAQ trend instead of weakening it. These are host-type shares, not the cascade classes, because duplication is tested first; they show how much the collisions could matter, not what §6 reports.

Provenance is decided from host and cross-host duplication, and never from a question's length, scaffolding, or type. Therefore, the form-to-provenance detector (§4) is not circular. However, one residual mechanism remains: manufactured-template is defined by cross-host duplication, and short strings duplicate more readily, so "length predicts manufactured" partly restates "short strings recur." We examine this by refitting on the `commerce-faq` subset, defined by host category rather than duplication. Here form barely separates the classes (AUC 0.554, against 0.725 for

the main fit), so dropping the duplication-defined template class takes most of the length signal with it. So the form-to-provenance detector separates genuine questions from template farms, not from commercial-FAQ language, and length carries what separation remains. We report this as the detector's boundary; the provenance and temporal results (§§5-6) do not depend on it.

C Form over time

The form classifier is deterministic regex and POS rules (one-shot, no training data) labelling three independent dimensions of a raw question string:

`type` a priority cascade, first match wins: `tag` (declarative with a ≤3-token tag after a comma, "…, right?") > `imperative-interrogative` (clause-leading base-form verb) > `wh` (a wh-tagged token) > `polar` (leading auxiliary or modal) > `elliptical` (no finite verb but a contentful noun/verb phrase) > `other`.

`person` `first` if a first-person pronoun (`i/we/my/our...`) appears, else `second` (`you/your`), else `impersonal`.

`scaffolding` present if the question spans more than one sentence, opens with a subordinator (`because`, `if`, `when`, …), or carries a politeness token (`please`, `thanks`, …); length is the content-token count.

Aggregated over the 2.2M sample, questions are 73.0% `wh` and 24.3% `polar` (`tag` 0.6%, `imperative` 1.0%, `elliptical` 0.2%, `other` 0.9%); 43.9% impersonal, 32.6% second-person, 23.6% first-person; 8.5% scaffolded; mean length 12.8 tokens. Figure 3 plots the two monotonic aggregate trends behind §6; the type distribution stays flat (`elliptical` a 0.2% floor).

Separated by provenance, every class shortens and de-scaffolds monotonically across the five eras *except* genuine-UGC, which stays flat (from 12.8 to 13.4 tokens; scaffolding from 7.0% to 8.5%, non-monotonic) while manufactured-template (from 8.6 to 7.8), commerce-FAQ (from 13.3 to 11.5), editorial (from 14.9 to 13.6), and unknown (from 14.2 to 12.8) all fall. So the corpus-level form decline is within-class degradation plus composition shift and not composition alone. Genuine-UGC's sampled base diminishes over the same window (from ∼38k to ∼6k occurrences), consistent with its decreasing share, so we read its trajectory as flat-within-noise rather than increasing.

Length dominates the §4 model, so we ask what remains once it is removed. On 60k questions per class (all 110 snapshots, spaCy tagger) we compute POS rates and lexical diversity, reweighting every class to the genuine-UGC length distribution over seven length bins so each comparison is made at equal length. Grammatically, the strata are close (manufactured vs. genuine at equal length: adjectives -4%, nouns -2%, verbs equal, content-word density -3%); they part on reference and repetition: proper nouns $+35\%$ (brand and product names; commerce-FAQ is the extreme at .084 vs. genuine .041), pronouns $+14\%$ (second-person calls to action), and type-token ratio (TTR) over a fixed 200k-token budget 11% lower. The template class's low TTR (.069 vs. genuine .090) partly restates its duplication-based definition. Commerce-FAQ, which is not as defined, is at .094, so we take the conservative reading. Manufactured questions are built much like genuine ones but name different things and repeat themselves more. Deeper syntactic and discourse analysis would require a parser pass that we did not run.

An external cross-check corroborates the manufactured form signature. WebFAQ (Dinzinger et al., 2025, 2026), mined from schema.org *FAQPage* markup, matches our manufactured class on every measure (8.96 tokens vs. our manufactured 8.5 and genuine 12.9; 4.0% scaffolded, 74.7% impersonal). This is evidence that markup mining harvests manufactured questions and treats them as natural asking. This is form-only, so WebFAQ's small-business host tail does not overlap our site-type labeler, so we do not compare provenance.

D Human validation

A human validation sample stratified by provenance class and era is labelled by three annotators: a shared 50-item calibration block for a three-way inter-annotator overlap, plus 100 disjoint items split across the three, 150 distinct items in all (∼83 per annotator). Each item is shown with the page's *as-crawled* text, so pages that have since died or been repurposed are judged from their original content. Annotators judge three cues per question *with its host in view*, and authenticity (genuine / manufactured / unclear) is computed from them:

community the host is a person-to-person venue (forum, board, comment thread, Q&A),

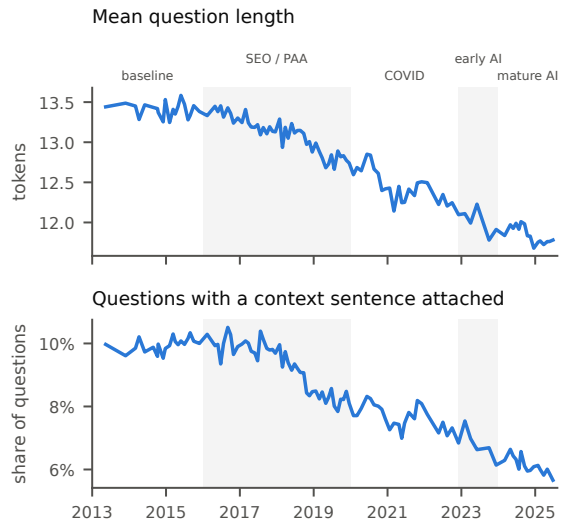


Figure 3: Form over time: mean length falls from 13.41 to 11.85 tokens (-11.6%) and the scaffolded share falls from 9.95% to 6.12% (-38.5%); both monotonic, MK $p<10^{-4}$. (Hamed-Rao $p<0.002$). Percentages are computed on unrounded era means.

judged from the host rather than whether the wording sounds human.

templated the exact string could recur across many sites (canonical FAQ, SEO or marketing boilerplate), which is a template signal, independent of length or person.

personal first-person, or about the asker's own situation.

A blind LLM rater (`gemini-3.5-flash`, run 2026-08) applies the same cues to each question's text, host, URL, and as-crawled page text, but never sees the human labels, never sets one, and only cross-checks them. The annotators are the authors, so this blind rater is the independent check on the labels, which is why we report inter-annotator κ and human-LLM agreement rather than treat the human labels as ground truth. Because provenance is decided from host and cross-host duplication, not from a question's form (App. B), the form-to-provenance detector is not circular; the human sample checks its target rather than restating it.

The human sample validates the detector's target: classifier provenance matches adjudicated human provenance at overall precision 0.73 (Wilson 95% CI $[0.66, 0.80]$), with genuine-UGC precision 1.00 (n=29); $[0.88, 1.00]$, so the lower-bound reading of the genuine share tolerates a true precision near 0.88, editorial 0.93, manufactured-template 0.80, and commerce-FAQ 0.74; the unknown residual is lower (0.43, n=49), as expected of a catch-all. Provenance inter-annotator κ is 0.52, which is moderate agreement (Landis and Koch, 1977). Per-class figures at n=150 are indicative since we read the intervals.

Era drift is the main issue when it comes to a twelve-year trend, so we also separate precision out by era (the sample is stratified on era, n=25-35 each). Excluding the unknowns, per-era precision is 0.81/0.95/0.90/0.83/0.93 from baseline to mature-AI, and genuine-UGC precision is 1.00 in every era (8/8, 6/6, 6/6, 5/5, 4/4). On the recall side, of the 37 items human annotators called genuine UGC the classifier caught 29 (0.78), lowest in the baseline era (8/11 vs. 4/5 in mature-AI), which is the direction that leaves the measured decline conservative, though at 5-11 items per era this is weak evidence and we treat it as such.

Authenticity itself (genuine / manufactured / unclear) is a harder, more subjective judgment: three-way κ is only 0.23. The disagreement is not random since agreement rises with how decisive the cues are: on the 50 shared items, annotators agree unanimously on authenticity for 60% of items where all three cues match (n=10), 46% (n=24), 21% (n=14), and 0% where none match (n=2). It also varies by host type, from 79% on genuine-UGC items to 25% on editorial, 14% on commerce-FAQ and 0% on the unknown class. Annotators split on the items whose host gives no clear signal, so the difficulty is in the construct and not the annotation. These counts are small, and even where all three cues match two in five items still split. Manufactured questions now read like genuine ones, which is the paper's central finding. What stays legible, and what the detector targets, is the host-level provenance these questions carry. We therefore base validation on provenance (which annotators *do* agree on) and report no form-to-authenticity refit. Form fields (type, person, scaffolding) were left almost entirely uncorrected against the classifier ($>$ 0.98 agreement); human-LLM agreement is likewise modest (κ=0.38, above the human-human level, and well below the agreement strong judges reach on easier constructs (Zheng et al., 2023)), an independent sign that the construct is the hard part. (The context rubric is validated by a single-annotator spot-check on a separate era-balanced sample, App. H.)

E Pre-specified hypotheses

The five directional hypotheses for the form-by-provenance test, fixed before the joint distribution was computed:

H1 (person) genuine-UGC is more first-person than manufactured. *Weak*: 29.7% vs. 21.4% (Cramér's V=0.04).

H2 (scaffolding) genuine-UGC is more scaffolded. *Held*: 7.3% vs. 3.9% (ϕ=0.07).

H3 (length) genuine-UGC is longer. *Held*: 12.9 vs. 8.5 tokens (d=0.71).

H4 (type) manufactured skews to templated/elliptical/impersonal forms, genuine to fuller wh-questions. *Failed*: question type carries almost no signal (V=0.03).

H5 (predictive) form features predict the genuine-vs-manufactured binary above chance. *Held*: AUC 0.725 (balanced class weights; 18%/82% class split).

F Multiple comparisons

We report trends for many curves, so we apply Benjamini-Hochberg (Benjamini and Hochberg, 1995) at q=0.05 across all 16 of them, on top of the Hamed-Rao correction each p-value already has. While fifteen survive, the one that does not is duplication at $D\geq10$ (p=0.061, q=0.061), which we already decline to claim; $D\geq5$ sits on the boundary (q=0.050). Ordered by corrected p, the classified provenance trends and the per-page rates clear the threshold comfortably (commerce-FAQ, the unknown class, manufactured-template per page, intensity and the two forum signals all at $q\leq0.004$), and the weakest surviving tests are genuine-UGC per crawled page and editorial per crawled page (q=0.021). Multiplicity therefore changes no claim in this paper. We note that this is a correction for testing many curves, and is independent of the crawl-size confound, which is what actually moves the duplication and manufactured-template share results (Limitations).

G The UGC host classifier

The genuine-UGC classifier is not a trained model: it is a deterministic function of the host string alone, with no training data and no threshold to tune, applied to the ∼1M most frequent hosts by occurrence; the long tail's UGC is caught by the URL patterns in the cascade (App. B).

A host is granted `genuine-UGC` only by (a) membership in a curated set of known person-to-person platforms, or (b) one of a small number of structural signals: a `forum`/`community`/`boards`-style subdomain, `forum` inside the registrable domain, or a forum-shaped URL path. Two rules make it precision-first. A bare `answers` or `ask` substring never grants genuine-UGC, since those are the strings SEO Q&A mills hide behind; such hosts route to a separate content-farm category. Anything else returns `unknown` rather than a speculative label, which is why the residual is large.

Because the classifier is deterministic over host strings, it has no era-dependent behaviour by construction. Its only channel for era-varying recall is drift in how hosts are named over twelve years. We cannot rule that out, and we do not have a host-level recall measurement stratified by era. What we have is the question-level check in App. D: recall on human-labelled genuine UGC is 0.78 overall and lowest in the baseline era (8/11, against 4/5 in mature-AI). At 5-11 items per era this is weak evidence, though the bias would understate the decline rather than inflate it. We release the classifier source, so the rules can be checked directly.

H Context-independence

Each question gets one of three labels by a deterministic POS-based rubric (spaCy tagger only; no parser or lemmatiser, so rules read token text, POS, and position), first match wins:

`conversational-fragment` second-person deixis (`you`/`your`), or a discourse-marker opener (`so`, `and`, `but`, `ok`, `well`) plus a comma, like someone mid-exchange.

`context-dependent` an unresolved reference outside the question: a dangling demonstrative (`this`/`that`/`these`/`those` as determiner or pronoun with no following head noun), an anaphoric `it` with no prior noun antecedent, or a definite reference to an unstated artifact (`the above`, `the following`, `either`).

`self-contained` none of the above.

Being POS-only, the rubric errs in both directions: it over-counts (expletive "it is true that…" reads as anaphoric) and under-counts open-class definite references ("the video", "both

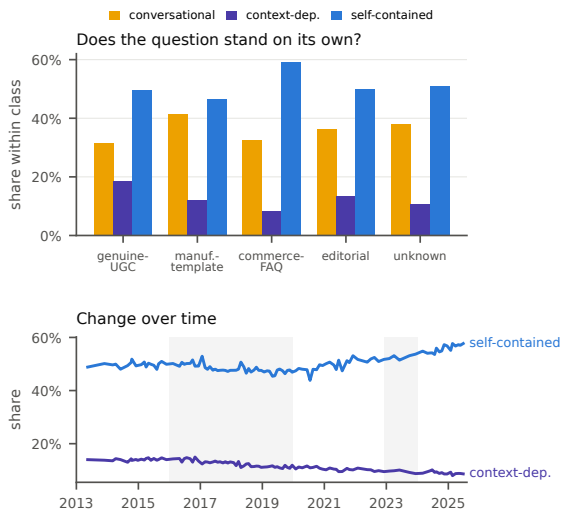


Figure 4: Context-independence of web questions. Top: label composition within each provenance class. Bottom: the context-dependent share falls over twelve years (significant under Hamed-Rao) while the self-contained rise is not (p=0.20)

documents") that fall outside its fixed `the above`/`the following`/`either` list. A blind, era-balanced single-annotator spot-check (n=60) performed by the author agrees with the rubric on the self-contained vs. needs-context split for 63% of items (the needs-context flag is 0.75-precise), so we read the per-class rates below as directional. Figure 4 gives the full within-class cross-tab and the two diachronic trends behind §7: genuine-UGC is the most context-dependent class (18.6%), commerce-FAQ the most self-contained (59.1%), and manufactured-template the most conversational (41.5%, second-person call-to-action boilerplate).

Wiktionary as a Crowdsourced Lexicon for English Dialects

Sidney Wong
Centre for Sustainability Research, University of Otago
Te Pūnaha Matatini Centre of Research Excellence for Complex Systems, New Zealand
sidney.wong@otago.ac.nz

Abstract

This paper evaluates Wiktionary as an ethically crowdsourced lexicon for English dialects. We took a two-phase approach, providing an in-depth descriptive analysis of the crowdsourced lexicon for 12 national varieties of English before applying the lexicon to geo-referenced, country-level social media language data to examine the real-world performance of this crowdsourced dialect lexicon. We demonstrate that Wiktionary matches or exceeds the coverage of traditional dictionaries, such as the Oxford English Dictionary (OED), for regional and Outer-Circle varieties. Our dialect-specific case study on New Zealand English found high alignment between Wiktionary and the OED based on word-formation patterns (R = 0.883). Similarly, we observed high alignment between the dialect lexicon and geo-referenced social media language. While this paper found that Wiktionary has broad coverage of lexical properties, it also highlighted some of the macro-challenges involved in evaluating dialect-responsive language resources and tools, such as the role of language contact in dialects and register effects in web-based corpora.

1 Background and Motivation

Despite recent advances in transformer-based large language models (Faisal and Anastasopoulos, 2025), an urgent need remains to develop dialect-responsive natural language processing (NLP) tools and resources. This is particularly challenging in low-resource language contexts, such as closely related varieties or dialects (Zampieri et al., 2020). A key reason for this limitation is the over-reliance on web corpora (including social media) during system development (Hovy and Prabhumoye, 2021). Consequently, for pluricentric languages such as English, French, Portuguese, and Spanish, contemporary systems remain prone to geographic bias (Faisal and Anastasopoulos, 2023).

With the popularisation of Web-as-Corpus approaches (Kilgarriff and Grefenstette, 2003), one approach to improving dialect responsiveness has been a shift towards more geographically representative corpora derived from social media language data (Dunn and Adams, 2020). This is because the metadata associated with these sources can be treated as ground truth for geographic dialect variation (Johnson et al., 2016). However, geographic information does not necessarily align with real-world linguistic behaviour (Sloan and Morgan, 2015). Other forms of geographic bias, such as non-local bias (Dunn et al., 2020), may further reduce representation within corpora.

Representative corpora are only one approach to improving fairness and equitable access in NLP systems. There is also a demand for resources and tools to evaluate dialect responsiveness. Dialect lexicons (including dictionaries and textual corpora) play a vital role in developing dialect-responsive language technologies (Joshi et al., 2025). For example, lexicons can be used to address the out-of-vocabulary (OOV) problem, especially in low-resource varieties or dialects (Joshi et al., 2025). With this in mind, we argue that ethically acquired open-source language resources, such as Wiktionary[1], provide essential ground-truth lexical information about dialects in pluricentric language contexts such as English.

1.1 Lexicons and Dictionaries

At the most foundational level, the lexicon comprises the basic building blocks of language. A lexicon is not restricted to words alone, but also includes affixes, stems, and more complex constructions such as phrasal units or idioms (Jackendoff, 2002). The creation and curation of lexicons remain core activities in dialectology (Chambers and Trudgill, 1998) and language documentation and

[1] https://www.wiktionary.org/

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 238–251
October 29, 2026 ©2026 Association for Computational Linguistics

conservation (Haviland, 2006). Some of the most well-known lexicons are dictionaries, which are reference works used for text reception, text production, or translation and were developed in response to the mass production of books and other written media (Bergenholtz, 2012). The most prominent example of a dictionary is the *Oxford English Dictionary* (OED), which, in its online form, amasses the meaning, history, and usage of over 500,000 words and phrases in English[2]. Institutional lexicons are by no means complete, with the primary challenges being size restrictions and the underspecification of lexical items (Klein, 2015).

Lexicons relating to non-standard' dialects (or varieties) may risk exclusion, as they are often omitted from dictionaries concerned with standard' language usage. Data collection efforts often rely on questionnaires and interviews with informants (Chambers and Trudgill, 1998). Naturally, the resulting lexicons can vary in size and quality. Using New Zealand English (NZE) as an example, the now-defunct New Zealand Dictionary Centre held a database of approximately 42,000 entries (Bardsley, 2009). However, this figure does not reflect actual usage. More contemporary sources, such as OED Online, include 1,404 entries coded to NZE alone. This is in sharp contrast to earlier sources, which combined New Zealand and Australian index entries into a lexicon of 820 entries (Turner, 1966). The exclusion of some lexical properties, such as *listemes* (Di Sciullo and Williams, 1987), also poses a challenge in lexicography because of the focus on individual lexical units (Klein, 2015).

1.2 Web-as-Corpus and Online Dictionaries

In the same way that the printing press played a transformational role in disseminating lexicons (Tarp, 2017), the internet has provided a new medium for the development of lexicons (Carr, 1997). As social media has become a major source of geo-referenced language data, computational approaches have increasingly been used to extract lexical dialect features (Nguyen, 2021). Grieve et al. (2017) documented the emergence of innovative lexical features on Twitter (now X) by measuring changes in relative frequency over time. Using a similar methodology, Mahler (2020) applied this approach to Reddit data. More recently, statistical language modelling has been used to extract lexical dialect features (Xie et al., 2024).

[2]https://www.oed.com/

Online dictionaries, such as Urban Dictionary and Wiktionary, have also provided another avenue for examining lexical information (Johns, 2019). Wiktionary is a multilingual, web-based dictionary collaboratively edited by users (Meyer and Gurevych, 2012). Similar to Wikipedia, Wiktionary is a form of digital commons. Unlike published dictionaries, which are curated by lexicographers, Wiktionary relies on registered volunteers to create individual entries. Users contribute the etymology, pronunciation, definitions, usage notes, derived forms, related terms, alternative forms, and translations for each entry. Wiktionary's crowdsourced nature also comes with limitations. Because volunteers can edit entries, their quality may vary, and they remain susceptible to errors and misinformation. In one notable example, nearly half of the articles in the Scots-language Wikipedia were authored by a non-Scots speaker, which raises serious concerns about the quality of crowdsourcing (Brooks and Hern, 2020).

1.3 Lexicons for NLP

Annotated corpora, lexicons, and grammars played a crucial role in the early development of NLP systems (Laporte, 2009). Lexicons provide not only morphosyntactic information but also syntactic-semantic information. One of the first lexicons for English, WordNet (Miller, 1995), provided semantic relations between words. Other lexicons providing syntactic-semantic properties of verbs include FrameNet (Fillmore and Atkins, 1994), VerbNet (Kipper et al., 2000), and COMLEX (Grishman et al., 1994). As a form of machine-readable structured data (Ylonen, 2022), Wiktionary is a valuable source of lexical-semantic information (Zesch et al., 2008). With the advent of vector-based word embedding models (Mikolov et al., 2013), Wiktionary-derived embeddings were used to adapt models to multilingual contexts (de Melo, 2015).

The demand for lexicons such as WordNet has reduced since the shift towards transformer-based LLMs (Meconi et al., 2025). Similarly, LLMs are now capable of producing fit-for-purpose dictionaries and lexicons for text reception, machine translation, and text production (de Schryver, 2023). However, this does not mean that the need for lexicons in the development of NLP systems no longer exists (Lew, 2024). As mentioned earlier, the OOV problem remains a significant issue in low-resource language contexts (Joshi et al., 2025). Even in state-of-the-art LLMs, a high concentra-

tion of OOV items can lead to a significant reduction in performance (Balde et al., 2025). Similarly, standard tokenisation and instruction tuning do not improve the performance of LLMs on reasoning tasks for 'non-standard' varieties of English (Lin et al., 2025).

1.4 Summary

The upkeep of labour-intensive dialect dictionaries and reference corpora is often unsustainable for low- or under-resourced language varieties. In the case of the New Zealand Dictionary Centre, operations ceased in 2012 (Bardsley, 2009). This is a major challenge as lexis is the most salient distinguishing feature of NZE (Deverson, 2000). While dialect lexicons may take years to curate (Klein, 2015), an advantage of Wiktionary as a source of lexical dialect information is its near-contemporaneous nature (Meyer and Gurevych, 2012). This is particularly important in the context of emergent internet-origin lexical dialect features (Mahler, 2020). Though there has been limited research on English dialects, Wiktionary has been applied to the dialects of other language contexts, including Germanic (Bernhard, 2014), Sinitic (Chang et al., 2022), and Romance languages (Sprugnoli et al., 2025).

The availability of an Application Programming Interface (API) enables the ethical acquisition of Wiktionary entries. This is particularly important for language tools and resources, such as digital corpora, to be consistent with FAIR (Findability, Accessibility, Interoperability, and Reusability) principles (Frey et al., 2020). With geographic dialect bias remaining a significant issue in the development of language models (Faisal and Anastasopoulos, 2025), there is a critical need to develop structured lexicons and dictionary-based features for underserved dialects to build fair and equitable NLP applications (Joshi et al., 2025). To evaluate the viability of Wiktionary as a crowdsourced dialect lexicon, this study addresses two central research questions:

RQ1 How does Wiktionary's regional English coverage compare to traditional lexicons such as the OED?

RQ2 Can Wiktionary-derived features identify regional dialect usage across social media text registers?

Category	n	Example
African	13	Botswanan, Cameroonian
Antarctic	-	-
Asian	3	East Asian, South Asian
Caribbean	8	Bahaman, Barbadian
Commonwealth	16	Australian, Bangaldeshi
European	6	British, Channel Islands
Middle Eastern	3	Egyptian, Israeli
North American	7	American, Bermudian
Oceanian	8	Fijian, French Polynesian

Table 1: Summary table of the Wiktionary categories of regional varieties including the number of subcategories (n) and examples.

2 Methodology

To address our research questions, we begin by comparing crowdsourced dialect lexicons from Wiktionary with existing published lexicons, namely the OED. We then evaluate these crowdsourced dialect lexicons against geo-referenced language data. Throughout our analysis, we contextualise the results and findings with reference to NZE. As one of the newest, most geographically isolated, and most well-studied Inner-Circle varieties of English (Hay et al., 2008), NZE is particularly well suited to exploring the efficacy of NLP tools and resources in a low-resource context. Although 96% of the NZE lexicon is shared with other Inner-Circle varieties (Deverson, 2000), lexis remains the most distinctive feature of NZE, characterised by a high degree of borrowings from te reo Māori as a result of language contact through settler colonialism.

2.1 Data Sources

One of the primary goals of this paper is to examine Wiktionary as an open-source dialect lexicon and as an alternative to institutional dictionaries, such as the Oxford English Dictionary (OED). In this section, we provide an overview of Wiktionary, as well as Reddit, which serves as our source of geo-referenced social media language data.

2.1.1 Wiktionary: The Free Dictionary

As of December 2024, the English edition of Wiktionary contained over 8.245 million entries. Of interest to the present study are the bespoke categories comprising entries associated with different language varieties. Entries are labelled with metadata that is then used to generate distinct categories. For English, entries are labelled as regional varieties and organised into broader geographic

Grouping	Country	Subreddit	Members	Submissions		Comments	
				n	$\bar{x}$	n	$\bar{x}$
Outer-Circle	India	r/india	3.2M	481K	64.5	12.5M	202.8
Outer-Circle	Pakistan	r/pakistan	575K	87K	56.9	2.4M	212.0
Outer-Circle	Malaysia	r/malaysia	1.3M	73K	55.3	3.5M	179.0
Outer-Circle	Philippines	r/Philippines	3.3M	331K	52.0	13.6M	138.7
Outer-Circle	Kenya	r/Kenya	219K	24K	42.3	475K	150.6
Outer-Circle	South Africa	r/southafrica	362K	56K	51.5	1.6M	212.2
Inner-Circle	Canada	r/canada	4.2M	175K	67.5	18.5M	241.3
Inner-Circle	United States	r/usa	119K	23K	54.8	102K	293.0
Inner-Circle	Australia	r/australia	2.7M	198K	60.0	12.9M	215.1
Inner-Circle	New Zealand	r/newzealand	794K	136K	56.2	8.7M	198.7
Inner-Circle	Ireland	r/ireland	1.2M	235K	53.3	10.5M	178.3
Inner-Circle	United Kingdom	r/unitedkingdom	5.4M	146K	67.3	7.4M	236.7

Table 2: Summary table of Reddit communities including the associated national variety of English (Country), the name of the community (Community), the number of subscribers as of June 2025 (Members), and the total number of words (n) and mean word length ($\bar{x}$) for Submission Posts and Comments.

categories (*Oceanian English*) and subcategories (*New Zealand English*). We provide a summary of these categories in Table 1. Some categories, such as *American English*, are further subdivided into subnational (*Midwestern English*, *Appalachian English*) or social (*African-American English*, *American slang*) varieties or dialects. Of interest to our analysis are the subcategories that correspond to national varieties of English.

National varieties can be classified as either Inner-Circle or Outer-Circle varieties (Kachru, 1985). The distinction between the Inner-Circle and Outer-Circle varieties reflects two distinct waves of British colonial expansion which shapes how these varieties are perceived, institutionalised, and used today (Kachru, 1985). The Inner Circle refers to varieties associated with countries where English is the majority language. Meanwhile, Outer-Circle varieties refer to those associated with former British colonies. We selected a geographically diverse sample of six Inner-Circle (Western Europe, North America, and Oceania) and six Outer-Circle (South Asia, Sub-Saharan Africa, and South East Asia) national varieties or dialects. One limiting factor was that this sample had to correspond with existing country-level communities on Reddit which we discuss in the next section.

2.1.2 Other Data Sources

In addition to Wiktionary entries, we use Reddit language data to validate our crowdsourced dialect lexicon. As a source of real-world dialect usage, we use implicit geographic information (such as place-based digital communities) as a proxy for geographic dialect variation (Hamre, 2024; Wong, 2026). Our primary data source comes from the Pushshift Dumps, an ongoing data collection effort that captures the top 40,000 Reddit communities via the Pushshift API (Baumgartner et al., 2020). We restrict our sample to country-level communities associated with twelve national varieties of English. These correspond with six Inner-Circle and six Outer-Circle varieties based on the Three Circles of English model (Kachru, 1985). Corpus characteristics are provided in Table 2.

One consideration is how Reddit functions as a register (Biber and Conrad, 2009). Reddit does not consist of a single text type but instead serves multiple communicative purposes. With this in mind, we compare two contrasting text types: Submission Posts and Comments. Users generally use Submission Posts to pose questions or seek advice from the community, whereas Comments mostly originate from community subscribers responding to posts. To illustrate the differences between these text types, the mean word count of Submission Posts is much lower ($\bar{x}$; Outer-Circle $= 58.1$, Inner-Circle $= 60.2$) than that of Comments ($\bar{x}$; Outer-Circle $= 175.0$, Inner-Circle $= 217.1$).

2.2 Analytical Pipeline

We take a two-phase approach to address the research questions. The results from Phase 1, Descriptive Analysis, are used to address RQ1, while the results from Phase 2, Corpus Validation, are used to address RQ2.

Phase 1: Descriptive Analysis To answer RQ1, we take a descriptive approach by comparing the coverage of lexical dialect features in the OED and Wiktionary. First, we compare the coverage rates of Wiktionary and the OED for each national dialect

variety. Second, we manually code the Wiktionary entries for NZE based on word-formation processes and cross-validate them against the OED. The Wiktionary entries for NZE offer a suitable dialect-specific case study because of their manageable size ($n = 1,415$). The coded information includes word-formation processes, Wiktionary definitions, stem or root words, morphological processes, usage notes, and other varieties that share the lexical feature.

Phase 2: Corpus Validation At this stage, we introduce the geo-referenced language data from Reddit. We use the same evaluation method as Grieve et al. (2017), comparing relative word frequencies. This is calculated as the rate of occurrence per 100,000 words (μ_i). This measure allows us to adjust for overall sample size and account for the varying sizes of each country-level community. More specifically, this phase enables us to evaluate and validate the effectiveness of the crowdsourced dialect lexicon using real-world language data.

2.3 Evaluation

The results from our analysis will be largely descriptive. As a relative measure of word frequency, μ_i allows us to directly compare the occurrence of lexical dialect features across corpora. As an example, *skux* ('stud, player, womaniser') should have the greatest μ_i in `r/newzealand` as feature that originated from New Zealand English. Although we do not conduct statistical modelling to test hypotheses, in Phase 2 we predict strong alignment between Wiktionary entries and subreddit communities and expect the highest mean relative frequencies of lexical dialect features to occur in their corresponding national subreddits.

Dialect	OED	Wiki	%Δ
Malaysian	19	450	2,268.4
Filipino	93	1,493	1,505.4
Canadian	550	2,834	415.3
Indian	721	1,816	151.9
Irish	1,145	2,105	83.8
Australian	2,435	3,381	38.9
New Zealand	1,435	1,415	-1.4
South African	982	928	-5.5
American	15,923	11,899	-25.3
British	16,191	11,283	-30.3
Kenyan	-	65	-
Pakistani	-	132	-

Table 3: Summary table comparing the total number of entries from the OED and Wiktionary (Wiki) by national variety (Dialect).

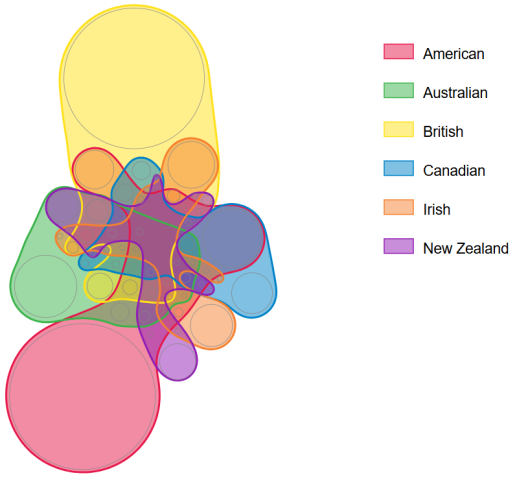


Figure 1: Euler diagram illustrating the degree of overlap in Wiktionary entries for the six Inner-Circle varieties of English. Similar to Venn diagrams, a Euler diagram visualises sets and their relationships particularly those with complex hierarchies.

2.4 Data Processing

An important consideration for crowdsourced corpora is data quality (Schäfer et al., 2013). We retrieved the entries as a `.json` file through the Wiktionary API using Python's `urllib.request` module. We then filtered out entries longer than three words, which approximately corresponds to trigrams. As an additional processing step, we removed duplicates by aggregating entries. We applied a similar processing procedure to the geo-referenced language data, first removing punctuation and special characters (excluding '&' and '$') from each observation string before converting all uppercase characters to lowercase. Of particular relevance to NZE is the variable use of diacritics to represent long vowels in te reo Māori orthography (Deverson, 1991): with a *tohutō* ('macron'), as two consecutive vowels, or with no diacritic. For the purposes of this paper, we did not standardise the Wiktionary entries. Once we had cleaned the observation strings, we joined them into a single string, which was then split into unigrams, bigrams, and trigrams using `nltk.ngrams()` from the `nltk` package. We then used the `Counter` class from the `collections` module to produce a dictionary object of key-value pairs, which we converted into a Pandas `DataFrame` for easier manipulation.

2.5 System Requirements

The analysis was carried out in Google Colab using Python with the Pandas, NumPy, and NLTK packages, and bar graphs were generated using Matplotlib. Euler diagrams were generated in R using the `eulerr` package.

2.6 Data Availability

The processed Reddit language data covering six Inner-Circle and six Outer-Circle varieties are available in the Hugging Face repository Reddit Corpus of Global Language Use (ReCGLU)[3]. All associated code and output datasets (in .csv format) are hosted in the paper's Hugging Face repository[4].

3 Results

A comparison of coverage as of December 2024 (Table 3) reveals that, while the OED remains slightly larger overall (by 13.1%; $n = 37,679$ vs. $n = 33,043$), the two resources capture distinct dimensions of Inner-Circle English. The OED retains higher coverage of American and British English (exceeding Wiktionary by 33.8% and 43.5%, respectively). In contrast, Wiktionary demonstrates superior coverage of regional Inner-Circle varieties, offering larger lexicons for Canadian, Irish, and Australian English. Regarding Outer-Circle Englishes, Wiktionary provides coverage for all six national varieties. By comparison, the OED contains entries for South African English, Indian English, Malaysian English, and Filipino English but lacks coverage of Kenyan English and Pakistani English. As shown in the Euler diagram in Figure 1, lexical sharing across Inner-Circle varieties aligns closely with regional and historical dialect clusters. The strongest cross-varietal overlap occurs between North American Englishes (American and Canadian, $n = 1,476$) and British and Irish Englishes ($n = 914$). In contrast, NZE exhibits a distinct Australasian/Oceanian alignment, sharing 400 entries exclusively with Australian English and a further 201 entries with both British and Australian English.

3.1 Dialect-Specific Case Study

We now investigate Wiktionary's coverage of one Inner-Circle variety: NZE. We found that borrowings from te reo Māori in Wiktionary were included in a separate category: *English terms borrowed from Māori*. This category contained 324 entries. The combined total of the two NZE-related categories was 1,541 entries, or 1,332 entries after applying the data-processing steps. When New Zealand English entries only, we observed a significant overlap with British and Australian entries in as shown in the Euler diagram (Figure 2). This was expected due to the historic relationship between the three varieties. Of the processed entries, only 498 were unique to NZE, with the majority being borrowings from te reo Māori ($n = 312$) and the remainder being formed through other word-formation processes ($n = 186$). We present a comparison between Wiktionary and the OED for lexical features classified according to their word-

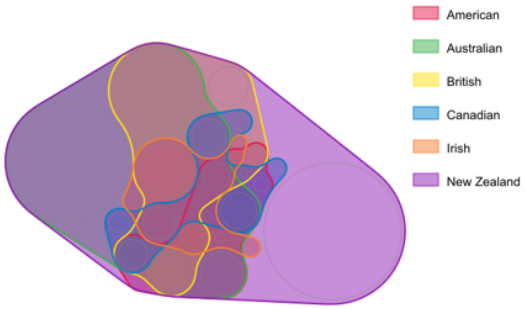


Figure 2: Euler diagram illustrating the degree of overlap in Wiktionary entries filtered by NZE and the six Inner-Circle varieties of English.

Word Formation Process	OED	Wiki	%Δ
Initialism	3	14	366.7
Blend	3	9	200.0
Backformation	3	6	100.0
Unknown	21	38	81.0
Proper name	37	66	78.4
Borrowing	287	399	39.0
Conversion	80	80	0.0
Derivative	230	228	-0.9
Imitative	8	6	-25.0
Variant	65	41	-36.9
Shortening	59	34	-42.4
Compound	587	335	-42.9
Uncertain	10	-	-
Arbitrary	4	-	-
Acronym	-	3	-
Idiomatic	-	73	-

Table 4: Summary table comparing the number of entries classified by word formation processes between the OED and Wiktionary (Wiki). The categories are ordered by percentage difference (%Δ) between OED and Wiktionary.

[3] https://doi.org/10.57967/hf/8197
[4] https://doi.org/10.57967/hf/10458

Dialect	Submission Posts			Comments		
	Community	μ_i	Alignment	Community	μ_i	Alignment
British	r/unitedkingdom	5,475.2	Yes	r/malaysia	3,973.3	No
Irish	r/ireland	889.1	Yes	r/ireland	771.1	Yes
Australian	r/australia	743.4	Yes	r/australia	333.4	Yes
New Zealand	r/newzealand	829.5	Yes	r/Philippines	498.8	No
American	r/usa	5,538.7	Yes	r/canada	3,582.3	No
Canadian	r/canada	443.2	Yes	r/canada	211.9	Yes

Table 5: Summary table presenting the Subreddit communities (Community) with the greatest mean relative frequency (μ_i) and alignment by Inner-Circle variety of English (Dialect) grouped by text-type: Submission Posts or Comments.

formation processes in the OED in Table 4. There is a strong positive correlation between the numbers of entries for each type of word-formation process (Pearson's $R = 0.883$; $R^2 = 0.780$) when we exclude the Uncertain, Arbitrary, Acronym, and Idiomatic categories. Furthermore, the number of entries from the additional category was comparable to the number of te reo Māori entries ($n = 337$) in the OED.

3.2 Corpus Validation

We present the mean relative frequency (μ_i) for each Inner-Circle variety of English and its prevalence in each country-level subreddit community in Table 5. For Submission Posts, we observed strong alignment between each country-level variety and its associated subreddit community. This result was expected. Even with a small subscriber base, we observed this relationship in r/usa. However, the relationship observed in Submission Posts did not hold for Comments. The only varieties of English that maintained a consistent relationship across both Submission Posts and Comments were Irish, Canadian, and Australian English. The prevalence of American English features in r/canada was expected, given that both are North American varieties of English. We visualised the mean relative frequencies (μ_i) for Submission Posts and Comments in the Appendix (see Figures 3 and 4).

4 Discussion

The findings suggest that Wiktionary complements traditional curated sources such as the OED and offers advantages in accessibility and coverage, particularly for Outer-Circle varieties and regional loanwords that are often neglected by Western-centric institutions. Based on this finding, we observe that the OED maintains broader historical coverage of primary Inner-Circle standard varieties (US and UK English), while Wiktionary matches or significantly exceeds the OED in capturing regional Inner-Circle varieties (Canadian, Irish, and Australian English) and Outer-Circle varieties (Filipino, Indian, and Malaysian English). Furthermore, Wiktionary's entry structure closely reflects lexicographical word-formation patterns ($R = 0.883$), while offering broader coverage of multi-word listemes and contemporaneous loanwords, as demonstrated by our dialect-specific case study.

Entry	Submissions	Comments
box of birds	-	11
box of fluffies	-	14
in the dogbox	-	7
jack a dandy	-	-
pack a sad	-	39
suck the kumara	1	5
turn to custard	-	36
up the boohai	-	5
warrant of fitness	5	278

Table 6: Table listing the total number of multi-word expressions (excluding borrowings from te reo Māori) in r/newzealand from Submission Posts and Comments.

Broad Coverage of Lexical Properties Coverage remains a limitation of dialect lexicons (Bergenholtz, 2012). When compared with historical dialect lexicons, only 121 entries appeared in Turner (1966), suggesting either that 85.24% ($n = 699$) were no longer in general usage or that the majority of those entries were more relevant to Australian English (of those 121, only 50 were distinctive to NZE). While the coverage of Wiktionary and the OED was comparable for most national dialects, the most prominent difference between them was Wiktionary's inclusion of idiomatic entries (e.g., *box of birds*, *good as gold*). Also known as complex compounds or listemes, these entries address a known limitation of institutional lexicons, as listemes are typically excluded or not treated as

Entry	r/newzealand		r/Philippines	
	n	μ_i	n	μ_i
mo	1,902	0.6	1,319,868	403.9
pa	3,338	1.1	1,174,769	359.5
para	372	0.1	499,040	152.7
eh	17,914	5.9	435,229	133.2
po	857	0.3	309,069	94.6

Table 7: Summary table including the token frequency (n) and mean relative frequency (μ_i) for the top five NZE Wiktionary entries grouped by Subreddit communities.

distinct headwords in dictionaries (Klein, 2015). In NZE, some of these listemes function alongside the stative verb *to be* (e.g., *to be a box of birds*, *to be good as gold*). We present the token frequencies of these multi-word expressions in Table 6. Of the nine entries, only two were not present in the OED (*suck the kumara*, *turn to custard*). Even though Wiktionary had more comprehensive coverage of idiomatic phrases in NZE, the OED better reflects the actual usage of these phrases.

Language Contact in Dialects The over-representation of NZE lexical items in the Comments of r/Philippines warrants deeper investigation. We inspected the top five NZE dialect features in the Comments of r/Philippines: *mo*, *pa*, *para*, *eh*, and *po*. The over-representation of these features points to the unique challenges posed by borrowings and loanwords in some dialect contexts. Both te reo Māori and Tagalog belong to the Austronesian language family, resulting in shared lexical roots and structural overlap. We compare the token frequencies (n) and mean relative frequencies (μ_i) of these entries in r/newzealand and r/Philippines in Table 7. The unexpectedly high relative frequency of NZE features in the Comments of r/Philippines appears to stem from orthographic homonyms shared between Tagalog particles and te reo Māori loanwords in NZE. The presence of borrowings and loanwords is closely linked to the challenge of multi-word expressions. As presented in Table 8, a number of multi-word expressions from te reo Māori have been borrowed into NZE. One example, *te reo* ('language'), has a high rate of occurrence in both Submission Posts ($\mu_i = 11.8$) and Comments ($\mu_i = 3.8$). Further exploration is required to determine the role of language contact in the development of dialect-responsive resources.

Entry	Submissions	Comments
biddi biddi	-	-
hei tiki	-	42
ka pai	2	619
kapa haka	8	415
kia ora	121	2,645
kohanga reo	-	90
mako shark	-	6
moho pereru	-	-
te ao māori	2	307
te reo	155	11,421
tena koe	1	58
tino rangatiratanga	3	571
whare wānanga	-	22

Table 8: Total number of multi-word expressions (borrowings from te reo Māori) in r/newzealand from Submission Posts and Comments.

Register Effects in Social Media Language Though not the primary focus of this paper, the divergence in lexical alignment between Submission Posts and Comments underscores the crucial role of register in geo-referenced corpus analysis. As a register, Submission Posts may prompt users to signal regional context through lexis within the constraints of a limited word count. This self-imposed behaviour may have contributed to the close alignment between national dialect features and their corresponding subreddits, particularly for unigrams. Conversely, Comments represent interactive, conversational discourse. In this register, short unigram tokens and cross-linguistic particles (such as the Tagalog homonyms that overlap with te reo Māori borrowings in NZE) dominate frequency counts. This pattern also holds for multi-word expressions, as observed in Tables 6 and 8, which exhibited higher token frequencies in Comments than in Submission Posts.

5 Conclusion

This work addresses the challenge of building geographically responsive NLP tools by benchmarking Wiktionary against the OED and evaluating its performance on real-world Reddit data. Overall, our empirical results yield two main insights for computational lexicography and corpus design that allow us to address our original research questions.

RQ1: How does Wiktionary's regional English coverage compare to traditional lexicons such as the OED? Across the six Inner-Circle varieties, the total sizes of the two lexicons are remarkably comparable: Wiktionary contains 33,043 entries, compared with the OED's 37,679 entries (a

13.1% difference). The OED is more comprehensive for American and British English; however, Wiktionary has greater coverage of all other varieties of English.

RQ2: Can Wiktionary-derived features identify regional dialect usage across social media text registers? Wiktionary-derived features can accurately identify regional dialect usage on social media, but only when register context is taken into account (e.g., prioritising titles over comment threads) and when text-preprocessing pipeline rules preserve character-level diacritics and filter out cross-linguistic homonyms.

Wiktionary serves as a viable, highly responsive complement to curated dictionaries for English varieties. Some of the challenges encountered in evaluating and validating this crowdsourced dialect lexicon illustrate broader issues associated with the use of digital corpora, particularly the influence of language contact in dialects and the effects of register within a corpus. We encourage the use of crowdsourced dialect lexicons to supplement existing corpora, but not to replace them. Future work should evaluate the performance of crowdsourced language tools and resources across different language registers in low-resource contexts. There is also scope to explore how crowdsourced dialect lexicons can be used to evaluate or fine-tune LLMs.

Limitations

We acknowledge that there are differences in the editorial processes between institutional lexicons, like the OED, and the crowdsourcing approach of Wiktionary. Most prominently, crowdsourcing relies on the bottom-up tagging of entries, which differs from top-down lexicographical categorisation. We therefore emphasise that the OED and Wiktionary are not direct like-for-like comparisons, but rather complementary resources. We also recognise that this paper is exploratory and relies heavily on descriptive analysis. As discussed in the conclusion, there is further scope to evaluate crowdsourced lexicons under experimental conditions. Additionally, we recognise that paralinguistic factors have not been fully addressed in the present study, an illustration of which was observed in short unigram tokens for New Zealand English.

Although Wiktionary lists entries such as *mo* ('moustache') and *po* ('a chamber pot') under NZE, these items overlap orthographically with high-frequency Tagalog words. In Tagalog, *mo* serves as a second-person singular possessive pronoun and *po* as a formal honorific. Conversely, in te reo Māori, these terms are written with diacritics as *mō* (a preposition indicating future possession or purpose) and *pō* ('night' or 'darkness'). In another example, the unaccented form *pa* functions in Tagalog as an aspectual marker ('still, yet') or an informal noun ('father'), whereas *pā* denotes a 'fortified settlement' in te reo Māori (although recorded without the tohutō in Wiktionary). Likewise, *para* designates the 'king fern' (Marattia salicina) in te reo Māori but functions in Tagalog as a preposition ('for, in order to'), interjection ('stop'), or verb particle ('seems'). These examples highlight the importance of accounting for cross-linguistic and multilingual contexts when using unstructured natural language data.

Finally, a key limitation of this study lies in the structural constraints of using country-level subreddits as proxies for geo-referenced language data. Membership in these place-based communities does not guarantee that subscribers physically reside in or are linguistically representative of the target speech communities. However, a similar limitation exists for other social media platforms, such as X (Sloan and Morgan, 2015), none of which are fully representative. Online communities present inherent demographic imbalances; for instance, `r/usa` holds a relatively small membership base (approximately 119,000) relative to the population of the United States, preventing it from serving as a fully representative sample of American English speakers. More broadly, these demographic and sampling biases are not unique to Reddit, but represent an inherent constraint of all web-derived corpora used in dialectology and natural language processing.

Ethics Statement

The rapid rise of proprietary large language models developed by 'frontier' AI laboratories has coincided with a notable decline in the digital commons (Stalder, 2026). Consequently, the primary purpose of this study is not to advocate for the over-exploitation of Wiktionary as one of the few remaining open, crowdsourced language resources. Nor does this paper suggest that the skilled labour of professional lexicographers can or should be outsourced to volunteer communities on open platforms. Rather, this paper illustrates the evolving

role of lexicography in the digital age, demonstrating how crowdsourced lexicons can actively support the development of fair, representative, and equitable language technologies (Lew, 2024).

Acknowledgements

This research originated during the author's doctoral studies at Toi Hangarau | Geospatial Research Institute, Te Whare Wānanga o Waitaha | University of Canterbury (UC). The author thanks supervisory committee members Dr Benjamin Adams (UC), Dr Jonathan Dunn (University of Illinois Urbana-Champaign), Prof. Jen Hay (UC), and Dr Kong Meng Liew (UC) for their guidance and feedback. This publication was supported by a Postdoctoral Fellowship from Te Pūnaha Matatini, New Zealand's Centre of Research Excellence for Complex Systems. The author also extends sincere thanks to the programme chairs, organising committee, and four anonymous reviewers for their constructive feedback on earlier versions of this manuscript.

References

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2025. Evaluation of LLMs in Medical Text Summarization: The Role of Vocabulary Adaptation in High OOV Settings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22989–23004, Vienna, Austria. Association for Computational Linguistics.

Dianne Bardsley. 2009. Lexicography in New Zealand. Technical report, New Zealand Dictionary Centre, Wellington.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839, Atlanta, GA. PKP Publishing Services Network.

Henning Bergenholtz. 2012. What is a dictionary? *Lexikos*, 22(1):20–30.

Delphine Bernhard. 2014. Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian. In *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, pages 23–29, Reykjavík, Iceland. Association for Computational Linguistics.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, United Kindgom.

Libby Brooks and Alex Hern. 2020. Shock an aw: US teenager wrote huge slice of Scots Wikipedia. *The Guardian*.

Michael Carr. 1997. Internet Dictionaries and Lexicography. *International Journal of Lexicography*, 10(3):209–230.

J. K. Chambers and Peter Trudgill. 1998. *Dialectology*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, United Kindgom.

Kalvin Chang, Chenxuan Cui, Youngmin Kim, and David R. Mortensen. 2022. WikiHan: A New Comparative Dataset for Chinese Languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3563–3569, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Gerard de Melo. 2015. Wiktionary-based word embeddings. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, FL. Association for Computational Linguistics.

Gilles-Maurice de Schryver. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4):355–387.

Tony Deverson. 1991. New Zealand English lexis: the Maori dimension. *English Today*, 7(2):18–25.

Tony Deverson. 2000. Handling New Zealand English lexis. In Alan Bell and Koenraad Kuiper, editors, *New Zealand English*, pages 23–39. Victoria University Press, Wellington, New Zealand.

Anne-Marie Di Sciullo and Edwin Williams. 1987. *On the definition of word*. Number 14 in Linguistic inquiry monographs. MIT Press, Cambridge, MA.

Jonathan Dunn and Ben Adams. 2020. Geographically-Balanced Gigaword Corpora for 50 Language Varieties. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2528–2536, Marseille, France. European Language Resources Association.

Jonathan Dunn, Tom Coupe, and Benjamin Adams. 2020. Measuring Linguistic Diversity During COVID-19. In *Proceedings of The Fourth Workshop on the Fourth Workshop on Natural Language Processing and Computational Social Science*.

Fahim Faisal and Antonios Anastasopoulos. 2023. Geographic and Geopolitical Biases of Language Models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163, Singapore, Singapore. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2025. Testing the Boundaries of LLMs: Dialectal and Language-Variety Tasks. In *Proceedings of the 12th*

Workshop on NLP for Similar Languages, Varieties and Dialects, pages 68–92, Abu Dhabi, UAE. Association for Computational Linguistics.

Charles J Fillmore and B T S Atkins. 1994. Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography. In B T S Atkins and A Zampolli, editors, *Computational Approaches to the Lexicon*, page 0. Oxford University Press.

Jennifer-Carmen Frey, Alexander König, Egon W. Stemle, Achille Falaise, Darja Fišer, and Harald Lüngen. 2020. The FAIR Index of CMC Corpora. In *CMC Corpora through the prism of digital humanities*, pages 127–144. L'Harmattan, Paris, France.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language & Linguistics*, 21(1):99–127.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. Comlex Syntax: building a computational lexicon. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, volume 1, pages 268–272, Stroudsbourg, PA. Association for Computational Linguistics.

Clayton Hamre. 2024. *A Geographic Analysis of Lexical Variation in North American English Using Reddit Corpora*. Doctoral, State University of New York at Buffalo, Buffalo, NY.

John B. Haviland. 2006. Documenting lexical knowledge. In Jost Gipper, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Essentials of language documentation*, number 178 in Trends in Linguistics. Studies and Monographs, pages 129–162. Mouton de Gruyter, Berlin, Germany.

Jennifer Hay, Margaret A. Maclagan, and Elizabeth Gordon. 2008. The Origins of New Zealand English. In *New Zealand English*, pages 84–94. Edinburgh University Press, Edinburgh, United Kingdom.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Ray Jackendoff. 2002. What's in the Lexicon? In Sieb Nooteboom, Fred Weerman, and Frank Wijnen, editors, *Storage and Computation in the Language Faculty*, pages 23–58. Springer Netherlands, Dordrecht.

Brendan T. Johns. 2019. Mining a Crowdsourced Dictionary to Understand Consistency and Preference in Word Meanings. *Frontiers in Psychology*, 10:268.

Isaac L. Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. The Geography and Importance of Localness in Geotagged Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 515–526, New York, NY. Association for Computing Machinery.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural Language Processing for Dialects of a Language: A Survey. *ACM Comput. Surv.*, 57(6):149:1–149:37.

Braj B. Kachru. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk and Henry G. Widdowson, editors, *English in the world: Teaching and learning the language and literatures*, page 11-30. Cambridge University Press, Cambridge, England.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *American Journal of Computational Linguistics*, 29(3):333–348.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press.

Wolfgang Klein. 2015. Lexicology and Lexicography. In James D. Wright, editor, *International Encyclopedia of the Social & Behavioral Sciences*, 1 edition, pages 938–942. Elsevier, Oxford, United Kingdom.

Eric Laporte. 2009. Lexicons and grammars for language processing: industrial or handcrafted products? In Letícia Marcondes Rezende, Bento Carlos Dias da Silva, and Juliana Bertucci Barbosa, editors, *Léxico e gramática: dos sentidos à construção da significação*, Trilhas Lingüísticas, 16, pages 51–84. Cultura acadêmica.

Robert Lew. 2024. Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1):426.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342, Vienna, Austria. Association for Computational Linguistics.

Hanna Mahler. 2020. Lexical Emergence on Reddit: An Analysis of Lexical Change on the "Front Page of the Internet". *Lexis. Journal in English Lexicology*, 2020(16).

Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavalle, and Roberto Navigli. 2025. Do Large Language Models Understand Word Senses? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33897–33916, Suzhou, China. Association for Computational Linguistics.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*. Oxford University Press.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*, Scottsdale, AZ.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Dong Nguyen. 2021. Dialect Variation on Social Media. In Marcos Zampieri and Preslav Nakov, editors, *Similar Languages, Varieties, and Dialects: A Computational Perspective*, Studies in Natural Language Processing, pages 204–218. Cambridge University Press, Cambridge, United Kingdom.

Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. 2013. The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In *Proceedings in the 8th Web as Corpus Workshop*, page pp.7, Lancaster, United Kingdom.

Luke Sloan and Jeffrey Morgan. 2015. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*, 10(11):e0142209.

Rachele Sprugnoli, Giovanni Moretti, Domenico Giuseppe Muscianisi, and Eleonora Litta. 2025. Ciallabacialla! Modeling and Linking a Regional Lexical Resource to Include Sicilian in the Semantic Web. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 1093–1101, Cagliari, Italy. Association for Computational Linguistics.

Felix Stalder. 2026. The Decline of the Digital Commons. *Theory, Culture & Society*, page 02632764261432877.

Sven Tarp. 2017. The concept of dictionary. In Pedro A. Fuertes-Olivera, editor, *The Routledge Handbook of Lexicography*. Routledge, London, United Kingdom.

George William Turner. 1966. *The English Language in Australia and New Zealand*. English Language Series. Longmans, London, United Kingdom.

Sidney Gig-Jan Wong. 2026. *Language, place, and social media : geographic dialect alignment in New Zealand.* Doctoral, University of Canterbury, Christchurch, New Zealand.

Roy Xie, Orevaoghene Ahia, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Extracting Lexical Features from Dialects via Interpretable Dialect Classifiers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 of *Short Papers*, pages 54–69, Mexico City, Mexico. Association for Computational Linguistics.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as Machine-Readable Structured Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. Association for Computational Linguistics.

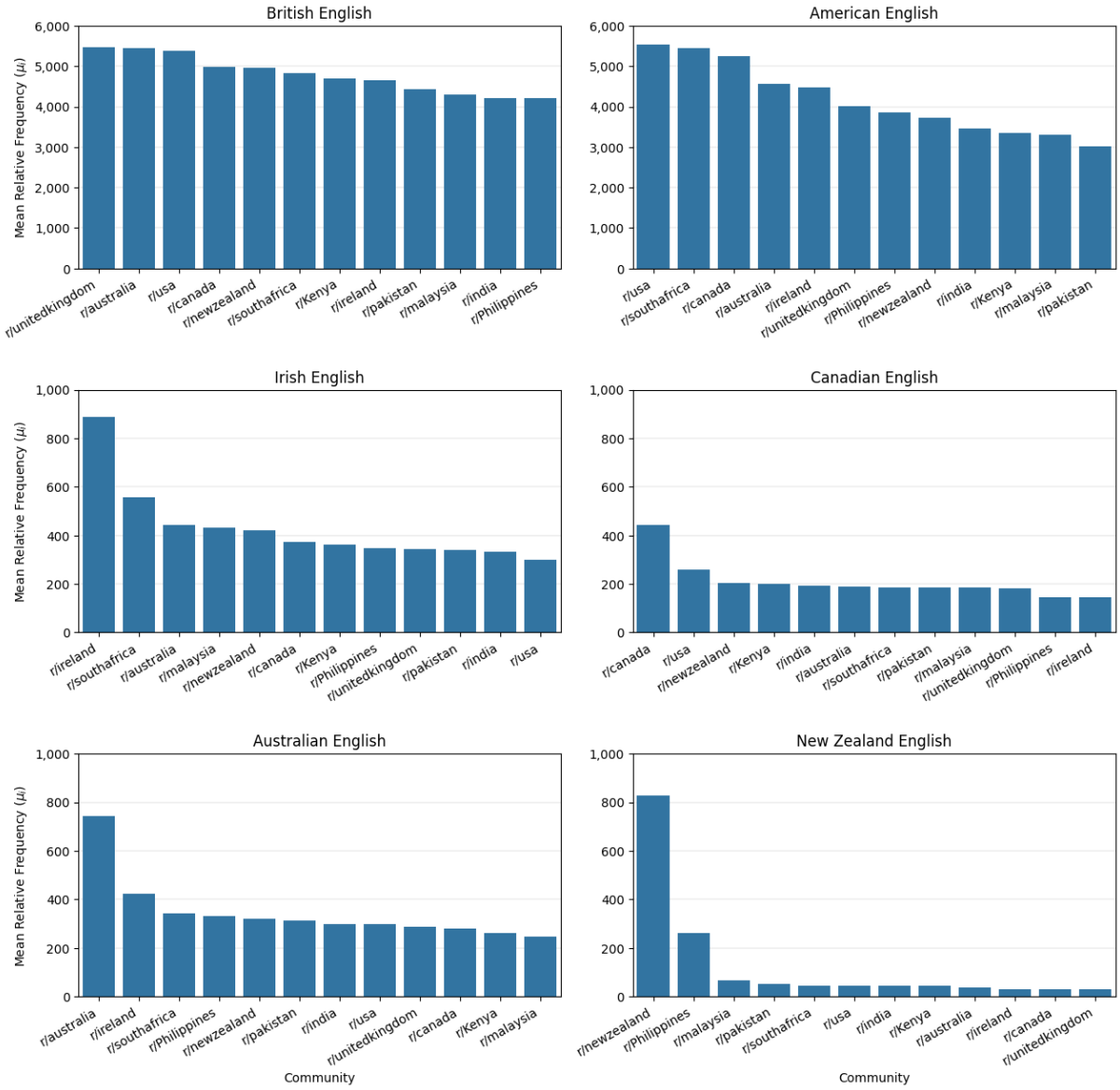


Figure 3: Relative frequencies of lexical dialect features associated with the six Inner-Circle varieties of English across twelve Reddit communities in Submission Posts.

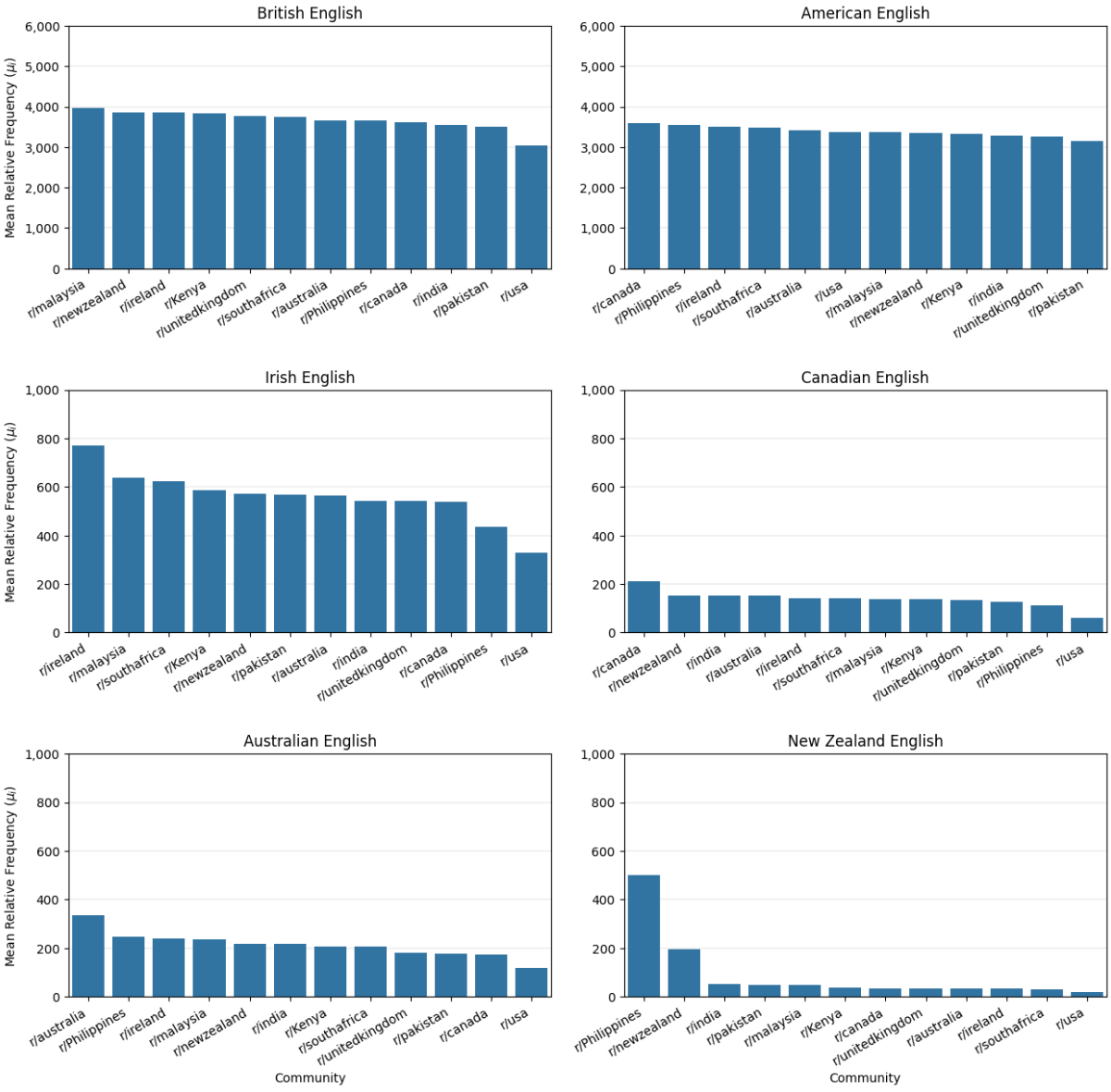


Figure 4: Relative frequencies of lexical dialect features associated with the six Inner-Circle varieties of English across twelve Reddit communities in Comments.

Task-Agnostic Web Document Annotation with LLM-Generated Descriptors

Otto Tarkka
University of Turku
TurkuNLP
ohitar@utu.fi

Erik Henriksson
University of Turku
TurkuNLP
erik.henriksson@utu.fi

Jenna Kanerva
University of Turku
TurkuNLP
jmnybl@utu.fi

Filip Ginter
University of Turku
TurkuNLP
figint@utu.fi

Abstract

Web-scale text corpora provide NLP researchers and corpus linguists with access to vast quantities of language data, but efficiently identifying the subsets of a large corpus that are relevant to a particular research aim remains a significant challenge. Existing approaches, including keyword search and task-specific classifiers, are often either too restrictive or depend on costly retraining for each new research question. We explore and evaluate the feasibility of a task-agnostic, LLM-driven descriptor schema for annotating web documents, using LLMs to generate natural-language labels that capture stylistic, tonal, topical, and structural properties of texts. Unlike prior work that emphasizes broad categorization or quality filtering, our approach yields fine-grained and interpretable descriptors designed to support flexible downstream applications. We evaluate the approach from multiple angles and show that descriptor-based annotation can enable flexible document retrieval while also exposing challenges related to schema harmonization and descriptor specificity.

1 Introduction

Natural language corpora have grown rapidly from the 1-million-word Brown corpus built in the 1960s to multi-trillion word corpora today. Large web corpora provide unprecedented insight into language use on the internet and are also used to train state-of-the-art Large Language Models (LLMs). However, large corpora are often only the starting point for a research project and are rarely used in their entirety. Corpus linguists typically only focus on a small subset of documents that are relevant to their research questions. Even for LLM training, requiring trillions of tokens, identifying the highest quality documents can be more important than focusing on quantity (Su et al., 2025; Lozhkov et al., 2024; Li et al., 2024; Diao et al., 2025).

To facilitate broader adoption of these vast collections in research, the data need rich annotations and metadata. In this paper, we build and evaluate a data-driven task-agnostic approach for generating open-ended document annotations. We leverage LLMs to generate a comprehensive document descriptor schema with more than 800,000 unique labels that can be used to create subsets of data for many purposes with relative ease. We present a proof-of-concept implementation of the approach and the steps taken in developing it and evaluate the descriptors for accuracy, aspect coverage, and usability. We release our code and the descriptor schema, which can be applied to any English corpus of web documents, and a ready-to-use, 10-billion-token descriptor-annotated corpus.

2 Background and related work

The idea of using the internet as a source of large-scale language data began at the start of the millennium with the Web-as-Corpus approach (Kilgarriff and Grefenstette, 2003). The value of metadata in large web corpora has also long been recognized (e.g., Barbaresi 2016) and, recently, LLMs have emerged as tools for annotating and analyzing text data at scale (Tan et al., 2024; Yadav et al., 2024; Chen et al., 2024a,b). ChatGPT and other state-of-the-art LLMs already outperform human coders in many annotation tasks (e.g., Bermejo et al. (2024); Gilardi et al. (2023); Törnberg (2024)). For example, Maragheh et al. (2023) use an LLM to generate thematic keywords for English news and e-commerce documents while Zhang and Takada (2025) use GPT-4o in an active learning setting, where documents are labeled with the LLM for classification tasks.

Identifying valuable documents for LLM training from large web corpora has become increasingly important (Penedo et al., 2024; Chen et al., 2024a; Su et al., 2025; Li et al., 2024). To this end, Wettig et al. (2025) introduce WebOrganizer, two parallel taxonomies that characterize the topic and format of web documents. They use Llama models

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 252–281
October 29, 2026 ©2026 Association for Computational Linguistics

to annotate documents and train smaller classifiers for subsequent large-scale annotation. Lozhkov et al. (2024) use an LLM to annotate web documents for their educational value and train a classifier to produce FineWeb-Edu. Idahl et al. (2026) use an LLM to annotate web documents along 18 quality axes such as completeness, structure, and information density.

All of the approaches above follow the same basic steps: (1) Annotate data. (2) Train a classifier on the annotated data. (3) Run inference on the full dataset. This requires considerable compute effort, as well as extensive human labor if the training data is manually annotated. These approaches rely on manually crafted, task-specific annotation schemas and categorize documents into a relatively small number of broadly defined categories with coarse labels. While this approach works, it limits the usability of the annotations to applications directly supported by the annotation schema. For instance, if one were to seek to retrieve "sarcastic" documents in the presently available large web corpora, no metadata in these corpora allow that query to be made. Consequently, the cycle of annotation, training, and inference on a large corpus must be repeated for this and every other new task not directly supported by the existing annotation schemas.

While training new classifiers is feasible for major undertakings, a myriad of smaller tasks and niche research questions remain unexplored due to this constraint. Therefore, in this paper we explore an approach where each document is given a task-agnostic annotation that is detailed enough to support a wide variety of tasks. In this bottom-up approach, an LLM generates a descriptor schema spanning hundreds of thousands of unique labels instead of relying on a few predetermined categories. Crucially, the data is annotated just once, after which the descriptors can be used with little cost and effort either on their own or as a first-pass filter. We propose our approach as a complement to, not a replacement for, traditional classification schemas. In our view, the more metadata large corpora have, the more accessible they are for all corpus users.

This philosophy of scalable, open-ended annotation shares conceptual ground with prior work, though with key distinctions. For example, TopicGPT (Pham et al., 2023), similarly employs an LLM to generate topic labels for documents. However, our method differs in several key respects: (1) TopicGPT assigns only topic labels, whereas our descriptors capture additional document aspects (see Section 4.2); (2) the scale of labels in TopicGPT is limited to dozens, rather than the hundreds of thousands in our schema; and (3) TopicGPT relies on predefined seed labels to guide annotation. Lu et al. (2024) also employ an LLM for open-ended annotation, but their work focuses on tagging LLM instructions to facilitate efficient model fine-tuning rather than general-purpose document description.

descriptor	
tag	**definition**
formal tone and language	The formal tone and language used convey credibility, authority, and expertise, making it suitable for professional or official contexts.

Figure 1: Example of a descriptor, consisting of a tag and a definition.

3 Method

Our objective is to assign each document a number of *descriptors*, where each descriptor captures some aspect of the document. Every descriptor consists of two parts: a unique *tag* and a *definition* (see Figure 1). The tag is a short phrase (usually 1–3 words) that holds the main meaning, while the definition expands on it by providing additional context and specificity. The definitions help distinguish between otherwise ambiguous or opaque tags, and add detail and length for better embeddings, which is useful for downstream tasks.

3.1 The descriptor schema

As the basis of our descriptor schema, we use a sample of 500,000 documents from FineWeb, a collection of relatively high-quality English documents (Penedo et al., 2024). This sample is selected to ensure comprehensive coverage of the diverse text types, topics, and formats present on the English web. We label the data using Llama-3.3-70B-Instruct (Grattafiori et al., 2024). In initial experiments, we found that this model generates high-quality descriptors while being computationally manageable at our data sizes.[1]

The annotation is carried out in an open-ended fashion, with descriptors created dynamically dur-

[1]While newer models have become available since our initial runs, re-running all experiments is unlikely to produce radically different results to justify the resource cost.

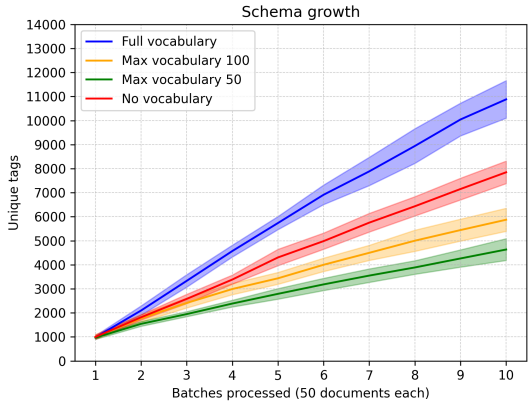


Figure 2: Rate of tag schema growth with different maximum vocabularies (average of 5 runs, batch size 50). Vocabulary size refers to the number of existing tags given in the prompt.

ing the annotation process. Since each document is labeled independently, a major methodological challenge is ensuring that the descriptors are detailed enough to cover all aspects of the document while avoiding semantic duplication, where the same phenomena are annotated with differently worded descriptors. The most straightforward approach would be to inform the model about descriptors assigned in previous data batches in the prompt and instruct it to introduce new descriptors only when necessary. Ideally, the annotation schema would grow quickly at first and then plateau, as most of the required labels would already exist. This approach would generate a comprehensive labeling schema in a single pass through the data. To examine this strategy, we conducted a preliminary experiment using tags alone, on the assumption that the model would find it easier to track existing tags than full descriptors. Figure 2 shows the growth of the tag vocabulary under three prompting conditions: providing the full, growing tag vocabulary, providing no vocabulary, or providing only the top N most common prior tags.

This experiment indicated that the model struggles to adhere to the given schema. We observed substantial levels of synonymous tags in the model output even when the model was informed of suitable existing tags. Moreover, contrary to expectations, providing a larger vocabulary accelerated the rate at which new tags were generated. Prompt tuning alone was insufficient to mitigate this issue.

Since generating the schema and labeling the documents in a single pass was found to be unfeasible, an alternative approach is to first generate descriptors freely and subsequently refine the schema by merging duplicates and synonyms. Our first attempt was to cluster and merge new descriptors with the existing schema after each new batch of data. While this slowed the growth of the descriptor schema, descriptors that were lexically similar but semantically opposed, such as "Formal" and "Informal", tended to get merged. To avoid this, the LLM was tasked with deciding whether and how to merge a cluster. While this approach worked better, throughput slowed drastically, preventing large-scale annotation. Hence, merging was implemented in two separate post-processing steps.

The final pipeline consists of three steps: (1) Generation: the initial descriptors are generated for each document independently; (2) Disambiguation: descriptors with duplicate tags are either merged, with their definitions revised accordingly, or renamed so that each unique descriptor has a unique tag; (3) Synonym merging: synonymous or near-synonymous descriptors are merged. These steps together produce a descriptor schema where each unique tag has exactly one definition and the prevalence of synonymous or near-synonymous descriptors is minimal.

To preserve space, the descriptions of the steps in the following sections are abbreviated. For a more comprehensive description and the full prompts used, refer to Appendices A and D.

3.1.1 Generation step

The generation step consists of three stages: generation, rewriting, and revision. First, the model is given one document and is prompted to "[d]escribe the document in aspects including, but not limited to, style, tone, genre, topic, domain, language, quality, sentiment, etc." For each tag, the model is also prompted to generate a definition, which should "explain its meaning and relevance to the document in 2 to 3 sentences". The generated descriptors should be detailed enough to support approximate reconstruction of the original content from the descriptors alone. However, to avoid mixing highly specific details with more general features, the model is prompted to place minute details under *specifics*. These are used to evaluate the model's ability to reconstruct the document but are not used thereafter.

To evaluate how well the meaning of the document is captured in the descriptors, we give the descriptors generated in the previous stage back to the model and ask it to rewrite the corresponding

document. The rationale behind this step is that a set of descriptors accurately reflecting as many aspects of a document as possible should, in principle, serve as a robust blueprint for rewriting it. If the generated text closely resembles the original, this suggests that the descriptors have successfully preserved its meaning.

Since the LLM might not generate high-quality descriptors in its first attempt, the model is asked to evaluate the quality of the rewrite by comparing it to the original and to revise the descriptors to produce a more faithful rewrite next time. The revision stage is followed by another rewrite, now using the revised descriptors, and then another revision. Although this process could be repeated arbitrarily many times, we stop after three sets of descriptors and rewrites to preserve computational resources.

To identify the set of descriptors that best captured the meaning of the original document, we embed the original document and each of its rewrites using the Stella-en-400M-v5 model (Zhang et al., 2025) and calculate their cosine similarity. The closer a rewrite is to the original document, the better the descriptors have summarized the document and thus, the better their quality. Our experiments confirm that rewrite similarity increases with more descriptor revisions (Figure 3).

Only the best descriptors for each document are kept. Labeling 500,000 documents results in nearly 7.5 million unique descriptors. Each document gets an average of 15 descriptors, and some of the most common tags are “formal tone”, “conversational tone”, and “technical”.

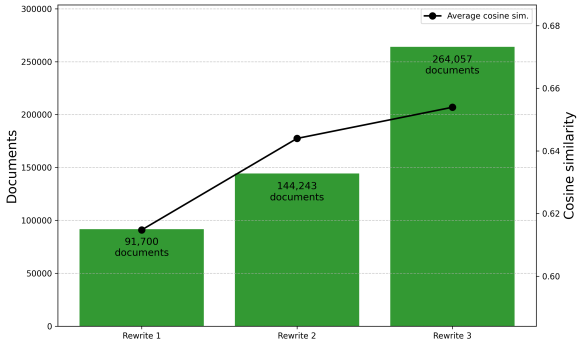


Figure 3: Number of documents whose best rewrite was generated on the first, second, and third attempts. The best rewrite was most often generated on the third attempt, although the average increase in cosine similarity was modest.

Step	Descriptors
Generation	7,433,036
Disambiguation	1,001,994
Synonym merging	800,488

Table 1: Number of unique descriptors after each step in the pipeline.

3.1.2 Disambiguation step

In the disambiguation step, we normalize all tags to lowercase and group together all definitions for a given tag. The vast majority of tags have only one definition, but some have up to thousands of different definitions. The groups with more than one definition are given to the LLM, which is prompted to either merge (near-)synonymous entries and provide one comprehensive descriptor or split the group into several descriptors, each consisting of a locally unique tag and definition.

3.1.3 Synonym merging step

After disambiguation, the schema still contains many (near-)synonymous descriptors because only exact tag duplicates were grouped together in the previous step. For instance, tags such as "formal tone" and "formal language" are still separate at this point. In the synonym merging step, each descriptor is first embedded using the Stella model, and mutually nearest neighbors are provided to the LLM. The model either accepts them as synonyms and chooses one to represent both or rejects their synonymity and keeps them as separate entries. After each iteration, new mutually nearest neighbors are identified and provided to the model until none are left. After synonym merging, the descriptor schema has condensed from nearly 7.5 million to around 0.8 million unique descriptors, where each tag is unique and has only one definition (Table 1).

3.2 Harmonization

The previous three steps, when carried out on a sufficiently large number of documents, will result in a schema of descriptors capable of describing a broad variety of documents. Any new documents can thereafter be annotated in this schema, without further updates. For new documents, we first apply the descriptor generation step (Section 3.1.1). This produces document-specific descriptors that do not necessarily correspond to the schema. We refer to these independently generated descriptors as *raw* descriptors throughout the remainder of the paper.

Next, these raw descriptors are mapped onto the

existing schema. We refer to the resulting schema-mapped descriptors as *harmonized* descriptors. For each raw descriptor, we retrieve the 10 most similar descriptors from the schema[2] and use the Llama model to select the best match among them. If the model rejects all 10 candidates, the descriptor is dropped, although in our experiments, this only happens to less than 0.5% of the descriptors. In the end, each document has a set of raw descriptors and a similar but schema-adhering set of harmonized descriptors.

While the unharmonized raw descriptors can be used as-is without harmonization, the harmonization step achieves two goals: First, it ensures the descriptors adhere to the frozen descriptor schema, which makes descriptor annotations between datasets and possibly different annotator models comparable and compatible with the same downstream processing methods. For example, examining the label distribution of two descriptor annotated corpora could reveal biases in their content. This comparison would be much more difficult to do without harmonized descriptors. Second, harmonization limits the descriptor space by reducing the number of unique descriptors. This is desirable because it makes the descriptors easier to use and greatly reduces the number of singleton descriptors (i.e., descriptors that appear in a single document), which leads to better recall in retrieval and classification tasks (see Sections 4.3 and 4.4).

4 Evaluation

We evaluate the descriptors on a sample of 100,000 FineWeb documents that were not used to derive the descriptor schema. We annotate the test documents with raw descriptors and then harmonize them and evaluate both raw and harmonized descriptors. On average, each document in the test set received 15 descriptors. Raw descriptors follow a long-tailed distribution, where the majority appear in only a single document, while the average harmonized descriptor appears in close to 5 documents (Figure 4).

For much of the evaluation, we rely on an LLM-as-judge approach, as human annotation would be prohibitively expensive. We use Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025) as the judge model. This model was chosen due to its strong performance and computational efficiency. It is from a different model family than the annotator model, which mitigates same-model bias. A small scale human annotation test is conducted to verify the reliability of the model (Section 4.1).

[2]Based on cosine similarity of their embeddings calculated using the Stella model.

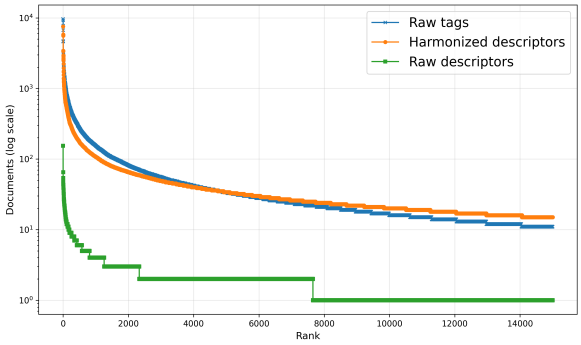


Figure 4: The number of documents in which the 15,000 most common tags and descriptors occur in the harmonized and raw settings. Harmonized tags not shown since their counts equal harmonized descriptors.

4.1 Evaluation 1: Descriptor accuracy

For the descriptors to be useful, they need to be accurate. While we can expect raw descriptors to be mostly accurate, since research has indicated that LLMs perform well as data annotators, some loss in accuracy is difficult to avoid during harmonization when a raw descriptor is mapped to its (near-)synonym from the schema.

We evaluate the accuracy of the descriptors on a sample of 5,000 documents from the test data. The judge model is given a document with its descriptors and is prompted to evaluate how accurately the descriptors describe the document on a five-point scale. The results are summarized in Table 2. Almost 90% of the raw descriptors are judged to be accurate, while only approximately 5% are inaccurate or mostly inaccurate. As expected, the harmonized descriptors are less accurate. Still, nearly 80% are judged to be at least partially accurate. This result shows that while the model does the initial descriptor annotation with high accuracy, the process of harmonizing the descriptors with an existing schema requires further development in future work.

We also have two human annotators annotate a sample of 20 documents and their descriptors along the same 5-point-scale as the LLM judge. The inter-annotator agreement (weighted Cohen's Kappa) between the two human annotators is 0.43 for raw descriptors and 0.58 for harmonized descriptors. Agreement between annotators and the LLM judge

Model judgment	Raw	Harmonized
Accurate	88.67%	61.81%
Mostly accurate	0.83%	2.90%
Partially accurate	5.58%	14.61%
Mostly inaccurate	0.55%	1.51%
Inaccurate	4.36%	19.16%

Table 2: Descriptor accuracy as determined by LLM judge.

ranges from 0.37 to 0.49 for raw descriptors and 0.43 to 0.49 for harmonized descriptors. These scores indicate moderate agreement between human annotators and the judge model. The modest kappa is partially explained by the skewed label distribution and resulting high chance agreement (see confusion matrices in Figures 6 and 7 in Appendix E). Most disagreements are adjacent (e.g., Accurate vs. Mostly Accurate), suggesting agreement on the general accuracy of descriptors but disagreement in fine-grained assessments. While the annotations are noisy, they provide reasonable confidence both in the accuracy of the descriptors and in the performance of the LLM judge model.

4.2 Evaluation 2: Document aspect coverage

A primary motivation for our method is to create metadata that can be used to create subsets of data for tasks that are not known when the annotation schema is created. Therefore, we analyze the aspects the descriptors cover and how they are distributed. We take a sample of roughly 5,000 documents, containing almost 75,000 descriptors, and use the same LLM judge model to categorize each descriptor into one or more aspect categories (see Appendix F for detailed description). Then, for each document, we measure if a given aspect category is present in at least one of its descriptors.

The results in Table 3 show that the most common aspect described by descriptors is Content & Subject Matter, followed by Style & Tone and Structure & Format. In terms of document coverage, we find that almost all documents have at least one descriptor relating to Content & Subject Matter. Coverage of other aspects is less common but still 69% of documents cover at least 4 aspects and 35% cover at least 5. The results for harmonized descriptors are very similar. In sum, document topics are the most covered aspect but other aspects are covered as well, including aspects not directly mentioned in the initial generation prompt.

4.3 Evaluation 3: Comparison to existing annotation schemes

In this section, we compare descriptors to two existing web document annotation schemes to assess how well the descriptors capture the information contained in these task-specific schemes. We apply the WebOrganizer classifiers, which assign each document with a format and a topic label (Wettig et al., 2025), to our test data. We also use the educational labels from FineWeb-Edu (Lozhkov et al., 2024). While these classification schemes provide a useful benchmark, machine-generated labels may contain inaccuracies. Subjectivity further complicates evaluation: we test this by using the LLM judge to evaluate whether the labels derived from the WebOrganizer classifiers are accurate. The LLM judge agrees with the classifier labels less than 60% of the time. Additionally, the descriptors offer a much more granular representation of content, with an average of 15 descriptors per document. With these limitations in mind, while we use WebOrganizer and FineWeb-Edu labels as comparative reference points, our results should be interpreted as indicative of broader trends rather than definitive ground truth.

First, we compare descriptors to WebOrganizer labels by providing the LLM judge with all the descriptors of a given document and asking which format and topic labels can be reasonably inferred from them. We find that 77% of topic labels could be inferred from both raw and harmonized descriptors, while inference rates for format labels were 57% (raw) and 53% (harmonized). This indicates that the descriptors often contain information about

Aspect	Descriptor coverage	Document coverage
Content & Subject Matter	43,284 (58%)	4,898 (99%)
Style & Tone	15,034 (20%)	4,304 (87%)
Structure & Format	7,236 (10%)	2,673 (54%)
Quality & Trustworthiness	3,837 (5%)	2,619 (53%)
Audience & Engagement	3,297 (4%)	2,600 (53%)
Ethical & Legal	1,164 (1%)	835 (17%)
Linguistic Features	1,103 (1%)	1,881 (38%)
Other	80 (<1%)	132 (3%)

Table 3: Aspect categories of raw descriptors according to the LLM judge. Descriptor coverage refers to the number of descriptors in the 5,000 document sample that cover the corresponding aspect. One descriptor may belong to more than one category. Document coverage refers to the number of documents that have at least one descriptor that covers the corresponding aspect.

a document's topic while information about its format is rarer, while still often present, which supports the results in Section 4.2.

Next, we train logistic regression classifiers using descriptors as features to predict topic, format, and educational labels and compare them against classifiers trained with TF-IDF features. This tests how much information relevant to predicting these labels is carried in the descriptors and how accessible this information is to a linear classifier compared with information encoded in n-grams. We split the test data into 80/20 training and test sets and target the binary FineWeb-Edu labels and the WebOrganizer topic and format taxonomies, each with 24 classes. For each task, we train a one-vs-rest logistic regression model.

We use the *tags* (not full descriptors) as features, encoded as a binary presence vector. In the harmonized setting, each tag has one definition, making harmonized tags and descriptors equivalent. In the raw setting, however, most full descriptors are unique to a single document, providing little basis for learning. Tags, in contrast, recur across documents and can thus serve as meaningful predictive features. The TF-IDF reference models are trained on 1–2-grams from full texts. The full per-class retrieval results appear in Appendix G (Tables 7–12).

The raw descriptors perform on par with TF-IDF capped at 50,000 features (Table 6) on macro-F1 on the educational (0.76 vs. 0.75) and format labels (0.48 vs. 0.49), but fall behind on topic (0.65 vs. 0.73). Even though our other evaluations show that topics are well represented in the descriptor space, their large number hints at very fine topic distinctions, which lead to sparse features and low performance. Raw descriptors outperform harmonized ones on all three tasks, reflecting the accuracy loss from harmonization.

The strength of descriptors as features is in their interpretability: for example, most of the highest-weighted features for non-educational documents are interpretable phrases ("product description" and "international travel guide"), whereas TF-IDF's are opaque tokens ("me", "my", and "2018") (Table 13, Appendix G).

We also test the applicability of the descriptor-trained models as cheap filters by ranking the documents by the model's score for each class to find the fraction of the test corpus that must be scanned to reach 80% and 20% recall. To reach 80% recall on topic, the descriptors require scanning 11.6–11.8% of the corpus against 7.2% for TF-IDF, and on format 22–24% against 16%. The gap narrows considerably at the top of the ranking: at 20% recall the descriptors scan nearly the same fraction as TF-IDF (1.3% vs. 1.1% on topic, 2.6–2.7% vs. 2.2% on format). Performance also varies by class: the descriptors are strongest on semantically coherent classes such as Adult, Religion, and Legal Notices, and weakest on classes defined by surface form or boilerplate, such as FAQ and Content Listing. Overall, the results in this section suggest that many of the distinctions made by existing purpose-built web annotation schemes are also reflected in the descriptors to some extent, despite the descriptors not being explicitly designed for them.

4.4 Evaluation 4: Retrieving documents without a classifier

One of our central goals is to enable fast and simple retrieval from a large corpus. To test this, we perform embedding-based similarity search for WebOrganizer topic and format labels from our test data. We embed each label together with its definition (Wettig et al., 2025, pp. 17–18) as a query. Additionally, we query for "sarcastic" documents as an example of a text type that could be of interest to researchers but is not covered by popular annotation schemes and is difficult to retrieve with a simple keyword or regular expression matching.

The evaluation is done in four steps: (1) Format each query as a descriptor (tag + definition). (2) Retrieve the nearest test set descriptors via Stella embeddings and a Faiss index (Douze et al., 2026), using a fixed squared Euclidean distance threshold of 320, which we experimentally find retrieves mostly plausible results without excluding too many potentially relevant hits. (3) Apply the LLM judge to filter retrieved descriptors for query relevance. (4) Use the resulting set of query-matched descriptors to retrieve the corresponding documents.

Using labels given by the WebOrganizer format and topic classifiers as the ground truth, we find that some labels are retrieved more accurately (e.g., Politics and Personal Blog) than others (e.g., Spam & Ads), while topic labels perform better than format labels (Table 4, full results in Appendix H). While raw descriptors have slightly higher precision, harmonized descriptors have better recall since a raw descriptor typically only retrieves one document, while one harmonized descriptor might appear in hundreds of documents.

In light of the issues with the WebOrganizer labels discussed in 4.3, we also evaluate the retrieval results with the LLM judge. The judge evaluates whether the documents associated with the approved descriptors correspond to the query. The judge is given three response options: "Yes","No", and "Partial", indicating full, no, and partial correspondence, respectively. The "Partial" option is given to reflect the more granular nature of the descriptors compared to the WebOrganizer labels, although the LLM judge rarely uses this option. This evaluation shows a higher rate of success for retrieved documents, with 56% and 75% full or partial correspondence for formats and topics, respectively, retrieved with harmonized descriptors. For formats and topics retrieved with raw descriptors, the model judges full or partial correspondence of 67% and 83%, respectively.

	Raw		Harmonized	
	Prec.	Rec.	Prec.	Rec.
formats	0.35	0.25	0.30	0.29
topics	0.64	0.35	0.57	0.42

Table 4: Average precision and recall for topic and format queries

The sarcasm query retrieved 5,459 documents with harmonized descriptors and 3,821 documents with raw descriptors. Since we do not have labels indicating sarcasm, we evaluate the retrieval of "sarcastic" documents only using the LLM judge. Out of the documents retrieved with harmonized descriptors, the judge model classified 74% as full or partial matches. For documents retrieved with raw descriptors, the acceptance rate is 85%.

As Figure 5 shows, descriptor acceptance generally declines with decreasing descriptor–query similarity. However, documents retrieved through LLM-approved descriptors are accepted at high rates, indicating that the approved descriptors are reliable indicators of sarcasm. The lower and more variable document-level acceptance rates for harmonized descriptors reflect their broader coverage: a single harmonized descriptor often retrieves many documents, some of which may not be relevant to the query.

This section has presented and evaluated a simple yet effective method to retrieve documents that would otherwise be difficult to identify, requiring only a user-provided query. While exhaustive recall is not guaranteed and the retrieved documents may include false positives, this method may serve as a fast way to obtain a substantial subset of relevant documents or as a preliminary filtering step to reduce the search space for more resource-intensive classification methods.

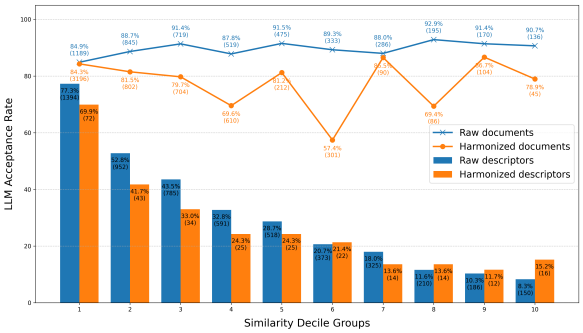


Figure 5: LLM acceptance rates for descriptors and documents retrieved by the sarcasm query, by descriptor–query similarity decile. Bars show the share of retrieved descriptors accepted by the LLM judge; lines show the share of associated documents accepted by the judge. Absolute counts are in parentheses. Document counts may exceed the number of unique retrieved documents because a document can be retrieved multiple times by different descriptors.

5 Discussion and conclusions

In this paper, we have explored a novel approach for enriching web corpora with metadata that is both human-readable and searchable while remaining sufficiently detailed to support a wide range of tasks without the need to specify them *a priori*. The main motivation of this study is to enable the categorization and retrieval of task-specific subsets of large web corpora without having to apply the resource-intensive annotate–train–classify approach separately to each new task. The effort of assigning the descriptors needs to be spent only once, after which they can be reused across applications. In our evaluation of the descriptor schema, we also implicitly evaluate the capability of LLMs for this task.

Our first finding is that LLMs can assign detailed, human-readable, and useful descriptors to documents, even without a predefined schema. First, we show that the descriptors accurately describe the documents (Section 4.1), although the current harmonization approach leads to a loss in accuracy. Second, we show that the descriptors cover many different aspects (Section 4.2). While topic is clearly the most covered aspect, style, tone, format, and quality are also represented, among others.

Third, we show that the descriptor schema contains much of the same information as two existing web document annotation schemes (Section 4.3). Finally, we show that a simple and fast query procedure can retrieve a non-trivial number of documents matching topics, formats, and even sarcasm; phenomena that would be difficult to identify in large corpora without a fine-tuned classifier (Section 4.4).

Our second finding is that while LLMs can generate detailed and useful descriptors, they struggle with systematicity across large document sets. Attempts to devise the descriptor schema and annotate the documents in a single pass were unsuccessful. We were able to alleviate but not fully solve the problem by disambiguating and merging the descriptors post hoc using a combination of embedding similarity and LLM judgments. Additionally, the high granularity of descriptors, while advantageous for the precise retrieval, complicates the identification of broad categories, such as "sports" or "news". Future work should explore new ways to enforce schema consistency and hierarchical descriptor structures for adjustable granularity and develop methods to efficiently use larger schemas. Some of our evaluation strategies, especially those in Sections 4.3 and 4.4 could be developed further into practical use cases.

To conclude, we strive to create rich, human-readable metadata for large corpora, produced only once and applicable to any number of tasks to complement existing annotation schemes. While the work presented in this paper is only one step in this direction and much work remains to be done, we envision the descriptors being used as a first-pass filter before more resource intensive classification or as a standalone tool for researchers lacking the means to train custom models. With further method development, such metadata could democratize access to large corpora for new user groups and increase the efficiency of current practices.

The pipeline and evaluation code are made available.[3] We also release a descriptor-annotated 10-billion-token sample of FineWeb, the full descriptor schema, the schema development and test sets as well as disambiguation and synonym merging traces.[4]

[3] https://github.com/TurkuNLP/LLM_document_descriptors

[4] https://huggingface.co/datasets/TurkuNLP/WebDocumentDescriptors

Limitations

This research is limited to monolingual English documents from the FineWeb dataset, which is purposefully built to exclude low-quality documents. While the schema is built and evaluated on separate subsets of FineWeb, it is not tested on fully out-out-distribution data and may not generalize well to noisier datasets. Additionally, the findings do not yet show whether the approach generalizes to other languages or multilingual settings. Future work should therefore evaluate the method on multilingual corpora.

In the current implementation, descriptors serve as document-level annotations. Although a given descriptor may apply to only a single paragraph or sentence within a document, the system does not explicitly track which specific section it pertains to. Future methodological refinements could incorporate section-level information into these annotations.

The specific language model used for descriptor generation and harmonization can also be a limitation. Since LLM capabilities continue to improve rapidly, newer models may produce more accurate, consistent, and multilingual descriptors. The exact prompts given to the model also have a big impact on model output. The prompts Consequently, the results reported here should be understood as reflecting the performance of one specific model rather than the upper bound of the proposed method.

Finally, although applying existing descriptors to downstream tasks is relatively fast and inexpensive, generating descriptors for large datasets remains computationally costly. To address this limitation, future work could explore smaller fine-tuned models that generate descriptors more efficiently while preserving descriptor quality. We plan to train and release such a model in future work.

Ethical considerations

This research relies on generative LLMs, which are known to produce errors in their output. Researchers should take care to evaluate the descriptor schema themselves before applying it to their research. While we use a high-quality dataset, FineWeb, as the basis of the descriptor schema, the documents might still contain biased, racist, sexist, or otherwise unreasonable or harmful content, which may be reproduced in the document descriptors. Using LLMs consumes electricity, which,

depending on the production method, can lead to greenhouse gas emissions. For this paper, processing 500,000 documents through the pipeline required an estimated 1,000 GPU hours on four AMD MI250X GPUs and all electricity came from renewable resources.

AI assistants were used to proofread sections of this paper. Some of the code used in the experiments was written with AI assistance. All AI-generated text and code were reviewed and verified by the authors.

Acknowledgments

This research has been supported by the Digital Europe Programme under grant agreement No. 101195233 (OpenEuroLLM); the Human Diversity consortium, under the Profi7 program (grant no. 352727) and the Centre of Excellence (grant no. 374221) by the Research Council of Finland; the Research Council of Finland through FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union – NextGenerationEU instrument), and the Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora (project 362459); and the GreenNLP project (project 353167) by the Research Council of Finland.

This project is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

We thank CSC – IT Center for Science, Finland, for computational resources.

References

Adrien Barbaresi. 2016. Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop*, pages 7–16, Berlin. Association for Computational Linguistics.

Vicente J. Bermejo, Andres Gago, Ramiro H. Gálvez, and Nicolás Harari. 2024. LLMs outperform outsourced human coders on complex textual analysis. SSRN Working Paper.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024a. Alpagasus: Training a better alpaca with fewer data.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024b. Is a large language model a good annotator for event extraction? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17772–17780.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan Su, Markus Kliegl, Zijia Chen, Peter Belcak, Yoshi Suhara, Hongxu Yin, Mostofa Patwary, Yingyan, Lin, Jan Kautz, and Pavlo Molchanov. 2025. Nemotron-CLIMB: CLustering-based Iterative Data Mixture Bootstrapping for Language Model Pre-training.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2026. The Faiss Library. *IEEE Transactions on Big Data*, 12(2):346–361.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar

Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng

Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models.

Maximilian Idahl, Benedikt Droste, Björn Plüster, and Jan Philipp Harries. 2026. propella-1: Multi-Property Document Annotation for LLM Data Curation at Scale.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29:333–347.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. 2024. DataComp-LM: In search of the next generation of training sets for language models.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. FineWeb-Edu: the Finest Collection of Educational Content.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024. #InsTag: Instruction tagging for analyzing supervised fine-tuning of large language models. The Twelfth International Conference on Learning Representations.

Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. LLM-TAKE: Theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4318–4324.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2023. TopicGPT: A Prompt-based Topic Modeling Framework.

Qwen Team. 2025. Qwen3 technical report.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-CC: Transforming common crawl into a refined long-horizon pretraining dataset.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Petter Törnberg. 2024. Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review*.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation.

Sachin Yadav, Tejaswi Choppa, and Dominik Schlechtweg. 2024. Towards automating text annotation: A case study on semantic proximity annotation using gpt-4.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models.

Yejian Zhang and Shingo Takada. 2025. Applying LLMs to active learning: Toward cost-efficient cross-task text classification without manually labeled data. *International Journal of Intelligent Systems*, 2025(1):6472544.

A Comprehensive description of the descriptor schema generation pipeline

A.1 Descriptor generation pipeline

We use the 70 billion parameter instruction fine-tuned version of Llama 3.3 to label our data. We find that this model generates high-quality responses while still being practical to run on large amounts of data.

The pipeline that creates the descriptor schema is divided into three steps. First comes the *Generation* step, in which the initial descriptors are generated. This step consists of three stages. After this comes

the *Disambiguation* step. During this step, most duplicate descriptors are merged into one or split into separate descriptors. Finally comes the *Synonym merging* step during which synonymous or near-synonymous descriptors are merged.

A.2 Generation step

The generation step is divided into three stages: initial generation, rewriting, and revision. In the initial generation stage, the model is given one document and is prompted to "[d]escribe the document in aspects including, but not limited to, style, tone, genre, topic, domain, language, quality, sentiment, etc." The model is asked to generate descriptors based on what it believes is essential for understanding, summarizing, or rewriting the document. The instructions are kept purposefully vague to not guide the model too much in a predetermined direction and to capture a broad range of descriptors. For each tag, the model is also prompted to generate a definition, which should "explain its meaning and relevance to the document in 2 to 3 sentences". The purpose of the definition is to provide additional context which is needed in later stages of the pipeline.

Since our aim is to build a general annotation schema, we do not want the descriptors to be overly specific. However in practice, we noted the model had a tendency to generate overly specific descriptors. For this reason, we offer the model the option to add more detailed descriptions into a list of "specifics". These are only used in the initial generation step but are discarded later.

A rewriting stage follows the initial generation. To evaluate how well the meaning of the document is captured in the descriptors, we give the descriptors and specifics generated in the previous stage back to the model and ask it to rewrite the document that they describe.

The rewriting stage is followed by a revision stage. After generating descriptors, specifics and a rewrite based on them, the model is asked to evaluate the quality of the rewrite and revise the descriptors accordingly. The model is prompted to compare the original document and its rewrite and to generate a revised set of descriptors that would better capture the meaning of the document.

The revision stage is followed by another rewrite stage, this time using the revised descriptors as input. This is followed by another revision stage and finally one more rewrite. Although this process could be repeated arbitrarily many times, we stop after three sets of descriptors and rewrites. To identify the set of descriptors that best captured the meaning of the original document, we embed the original document and each of its rewrites using the Stella-en-400M-v5 model Zhang et al. (2025), which is one of the best performing models in its size for English text clustering on the MTEB leaderboard (Muennighoff et al., 2023). The cosine similarity between each rewrite and the original is used as a measure of how much the rewrite resembles the original. The closer a rewrite is to the original document, the better the descriptors have summarized the document and thus, the better their quality.

After these steps, for each document, we only keep the descriptors that produced the best rewrite (measured by cosine similarity) and discard the rest. The specifics are also discarded at this stage.

A.3 Disambiguation step

In the disambiguation stage, we normalize all tags to lowercase and group together all definitions for a given tag. The vast majority of tags have only one definition, but some have more, with the most common ones such as "informal tone" having thousands of different definitions. The tags with more than one definition are provided to the LLM, which is prompted to either merge together synonymous or near-synonymous entries and provide one holistic descriptor or split the group into several descriptors, each with a locally unique tag and definition. To avoid overwhelming the model with extremely long inputs, large groups are split into subgroups of 20 definitions or fewer, each of which is disambiguated separately. As the model is not aware of decisions made in different instances, it can generate new, locally unique descriptors that are still duplicates globally. Therefore, this process is repeated several times until there are only a handful of duplicates left.

After 20 iterations, only four tags remain with more than one definition. The process is stopped here, since the model keeps inventing descriptors that already exists elsewhere in the data and the rate of lowering the number of duplicates approaches zero.

A.4 Synonym merging step

After disambiguation, there are still many synonymous and near-synonymous descriptors in the set because in the disambiguation step only exact tag duplicates were grouped together. For instance,

tags such as “formal tone” and “formal language” are still separate, despite being arguably synonymous in most contexts. In the synonym merging step, each descriptor is embedded with the Stella-en-400M-v5 model and mutually nearest neighbors are provided to the LLM. The model is prompted to either keep them as separate entries or accept them as synonyms and choose one as the representative, while the other is discarded. The discarded descriptor is thus effectively merged into the other. After each iteration, new mutually nearest neighbors are identified and provided to the model. The process ends when no more merges are made.

The synonym merging step merges most of the remaining duplicate tags that survived the disambiguation step, but leaves just one tag with more than one definition. To ensure that each tag has exactly one definition, we do a forced merge, where the remaining duplicates are provided to the model, which chooses one as the representative descriptor.

B Example document and descriptors

Document

- I am now the proud owner of a beautiful Black (the fastest colour of course) GCSRT8 here in Western Australia. There is a growing number of these beasts here although I have yet to see one on the road. You seem to have a fantastic SRT8 community thing going in this forum, I only wish that there was a GCSRT8 group like you here. (If there is an active SRT8 forum/group in Oz, can someone let me have the details?) I have gathered a lot of useful info from the forum to help me with my next move regarding performance mods and have decided the first will be to change my tstat to a 180 (and maybe the fan mod - although I am not that comfortable in cutting wires etc to get it done) Can anyone please advise where I can get these thermostats in Australia? What brands are available here? Part numbers? Our local auto chain store (Supa Cheap) hardly even knows what they are let alone if they stock them! Any help would be appreciated.

Document rewrite

- G’day mates! I’m a car enthusiast from Western Australia, and I’m stoked to be part of this awesome community. I’ve got a GCSRT8, which I’ve modified with an SRT8 kit, and I’m always on the lookout for ways to improve its performance. Recently, I picked up a 180 thermostat from Supa Cheap, and I’m thinking of doing a fan mod to keep things cool under the bonnet. I’ve heard it’s a bit of a tricky job, but I’m keen to give it a go. Has anyone else out there done this mod? I’d love to get some advice from you blokes with some technical expertise. I’ve been doing some research, and I’ve got the part numbers written down, but I’m not sure about cutting wires and whatnot. I’ve been to the local auto chain store, and they didn’t seem to know much about performance modification, so I thought I’d come to the car community for some help. By the way, what’s the fastest colour for a GCSRT8? Just kidding, mate! But seriously, thanks in advance for any tips or advice you can offer. I’m really appreciative of all the knowledge and enthusiasm on this forum, and I’m excited to hear from you all.

Raw descriptors

- car enthusiast; The document is written by a car enthusiast, indicating a strong interest in vehicles and their modification. The author’s passion for cars is evident in their language and tone.

- technical; The document discusses technical aspects of car modification, such as thermostats and fan mods. This indicates a focus on the mechanical details of the vehicle.

- Australian; The document is set in Western Australia, and the author is seeking information specific to their location. This context is important for understanding the author’s needs and concerns.

- informal; The document has a casual and friendly tone, indicating that the author is approachable and enthusiastic about the topic. The use of colloquial language and expressions conveys a sense of camaraderie.

- modification-focused; The document discusses plans for modifying the vehicle’s performance, including changing the thermostat and potentially installing a fan mod. This indicates a focus on improving the vehicle’s capabilities.

- seeking advice; The author is seeking guidance and recommendations from the community, demonstrating a willingness to learn and take advice from others. This shows that the author values the opinions and expertise of others.
- personal; The author is a proud owner of a GC-SRT8 and is seeking to modify and improve their vehicle, indicating a personal investment in the topic. The author's ownership and enthusiasm are central to the document.
- regional; The author is seeking information specific to Australia, including availability of parts and local resources. This context is crucial for understanding the author's needs and concerns.
- appreciative; The author expresses gratitude for the community's potential help and advice, demonstrating a positive and respectful attitude. This tone is important for building relationships and trust within the community.
- enthusiastic; The author is excited about their vehicle and the potential modifications, conveying a sense of eagerness and anticipation.
- community-oriented; The author is seeking to connect with others who share their interests and is looking for a sense of community and belonging.

Harmonized descriptors

- automotive enthusiast; The author is an automotive enthusiast with a strong interest in modifying and improving vehicles, possessing deep knowledge of cars and trucks, and valuing high-quality service.
- automotive performance; The document focuses on the automotive domain, discussing vehicle performance, modifications, and technical aspects.
- western australia context; The text is specific to Western Australia, highlighting the regional focus of the document.
- informal social; The tone of the document is casual and friendly, indicating a personal or social relationship between the writer and the reader, and describes a situation that occurs in a social setting, making it interesting and engaging, and involves colleagues in a lighthearted activity.
- performance mods; The document discusses various modifications made to vehicles to improve their performance, including upgrades to engines, transmissions, suspensions, and other parts, as well as customization and DIY efforts.
- user seeking advice; This descriptor captures the purpose of the post, which is for the user to seek information or advice from others.
- passion and identity; The author's passion for cars is a central aspect of their identity and is reflected in their language and experiences, with a focus on modifying and customizing their vehicle.
- relevance to australia; The document is specific to the Australian context, making it relevant to an Australian audience, with references to local policies, laws, and regulations.
- polite and appreciative tone; The tone is polite and appreciative, with the authors expressing gratitude for any help they might receive.
- enthusiastic mood; The tone is positive and enthusiastic, conveying excitement and anticipation.
- online community seeker; The author is looking for a community where they can express themselves freely and engage in discussions with others who share similar interests. This descriptor emphasizes the author's desire for connection and dialogue.

Specifics

- GCSRT8
- SRT8
- Western Australia
- Supa Cheap
- 180 thermostat
- fan mod
- part numbers

- performance modification
- car community
- technical expertise
- fastest colour
- cutting wires
- forum
- local auto chain store

C Query phrases

The exact phrases used to query documents. The phrasing of the topic and format queries are derived from the WebOrganizer paper.

- **Formats**
  - About (Org.); The document is formatted as an organizational About page, typically containing a self-description or introduction by an organization such as a company, university, government agency, or non-profit, but the content is not necessarily verified and may contain self-promotion
  - About (Personal); The document is formatted as a personal About page, typically containing a self-description, introduction, or profile information on a personal or hobby website
  - Academic Writing; The document is formatted as academic writing, such as a research paper, paper abstract, thesis, or literature review
  - Audio Transcript; The document is formatted as an audio transcript, providing a written record of spoken language, such as an interview, court hearing, movie, podcast, lecture, or speech
  - Comment Section; The document is formatted as a comment section or discussion forum with multiple posts or comments, such as a Reddit thread or a comment section on a news article or blog
  - Content Listing; The document is formatted as a content listing, providing an overview of content for navigation, such as a sitemap, product catalog, search results page, or news listing with short article snippets
  - Creative Writing; The document is formatted as creative writing, such as a short story, novel chapter, poem, or song lyrics
  - Documentation; The document is formatted as documentation, such as technical writing, API documentation, README files, or source code, intended for developers or experts rather than end-users
  - FAQ; The document is formatted as a Frequently Asked Questions page
  - Knowledge Article; The document is formatted as a knowledge article, written in an objective and neutral style and published on a moderated platform or by a reputable source
  - Legal Notices; The document is formatted as legal notices, such as terms of service, legal disclaimers, privacy policies, or license agreements
  - Listicle; The document is formatted as a listicle, presenting article or blog content in the form of a list, such as a "Top 10" list or "4 best places to visit in X", but excluding navigational lists of site content
  - News (Org.); The document is formatted as organizational news or announcements, such as a press release or blog post by a company, university, government agency, or non-profit organization
  - News Article; The document is formatted as a news article, written by journalists about current events and published by a news organization, but excluding long reads, profiles, editorials, journalistic essays, and newspaper interviews
  - Nonfiction Writing; The document is formatted as nonfiction writing, such as a long read, profile, editorial, essay, obituary, memoir, or other professional nonfiction text
  - Personal Blog; The document is formatted as a personal blog, typically written by an individual and relating personal experiences or opinions
  - Product Page; The document is formatted as a product page, typically containing descriptions or promotions for a product or service, including products in a wider sense such as university course descriptions

  - Q&A Forum; The document is formatted as a Q&A forum, with an explicit question-and-answer structure, such as Quora or Stack Exchange
  - Spam / Ads; The document is formatted as spam or advertising, consisting primarily of spam content, SEO keyword stuffing, short online ads, or content with no apparent purpose
  - Structured Data; The document is formatted as structured data, containing multiple data entries with a common structure, such as a table, datasheet, movie database, glossary, dictionary, JSON file, CSV file, or XML file
  - Customer Support; The document is formatted as customer support content, produced by an organization for a general audience, such as a troubleshooting guide
  - Truncated; The document is formatted as truncated or incomplete content, such as a paywalled page, login-gated page, incomplete page, video description, or image gallery, but excluding pages with multiple snippets of truncated articles
  - Tutorial; The document is formatted as a tutorial or how-to guide, such as a cooking recipe, DIY instructions, WikiHow page, or Khan Academy course, and must contain the actual tutorial content
  - User Review; The document is formatted as a user review, such as a review posted by users on Yelp, TripAdvisor, or similar platforms

- **Topics**
  - Adult; The document's topic is adult content, including porn and other explicit content
  - Art & Design; The document's topic is art and design, including architecture
  - Crime & Law; The document's topic is crime and law, including law enforcement (excluding financial crime, litigation, social issues and the legislative process)
  - Education & Jobs; The document's topic is education and jobs, including pedagogy, training & certification, and academia
  - Entertainment; The document's topic is entertainment, including music, movies, TV shows, videos, celebrities, humor, nightlife but excluding music or film if discussed as art
  - Fashion & Beauty; The document's topic is fashion and beauty, including clothing, accessories, and cosmetics
  - Finance & Business; The document's topic is finance and business, including taxes, regulations, investments, insurance, credit cards, personal finance, corporate communication, marketing, and human resources
  - Food & Dining; The document's topic is food and dining, including recipes, groceries, beverages, and restaurants, but excluding nutritional sciences
  - Games; The document's topic is games, including video games, board games, and gambling
  - Hardware; The document's topic is hardware, including computer hardware, phones, televisions, and other consumer electronics
  - Health; The document's topic is health, including medicine, wellness, mental health, veterinary science, and nutritional science, but excluding health insurance
  - History; The document's topic is history, including geography and archaeology
  - Home & Hobbies; The document's topic is home and hobbies, including real estate, renting, relocation, furniture, appliances, home improvement, DIY, gardening, pets, toys, and collecting
  - Industrial; The document's topic is industrial activity, including mining, agriculture, manufacturing, utilities, construction, raw materials, industrial goods, chemicals, and textiles, but excluding general business topics and business finance
  - Literature; The document's topic is literature, including literary criticism, linguistics, philosophy, and related subjects in the humanities, but excluding texts whose literary style is incidental to their topic

  - Politics; The document's topic is politics, including social issues, political campaigns, the legislative process, geopolitics, protests, and activism
  - Religion; The document's topic is religion, including spirituality
  - Science & Technology; The document's topic is science and technology, including physics, chemistry, biology, environmental science, mathematics, statistics, biotech, and engineering
  - Social Life; The document's topic is social life, including family, friends, relationships, and community, but excluding specific social activities that fall under other topics
  - Software; The document's topic is software, including the use of software and the internet
  - Software Development; The document's topic is software development, including algorithms, coding, and web development
  - Sports & Fitness; The document's topic is sports and fitness, including martial arts, motor sports, outdoor activities, and sports equipment
  - Transportation; The document's topic is transportation, including cars and other vehicles, taxis, public transportation, traffic, commuting, aviation, rail, shipping, and logistics
  - Travel; The document's topic is travel, including hospitality, hotels, sight-seeing, and cruises, but excluding detailed descriptions of tourist destinations

- **Sarcasm**
  - sarcasm; the document contains sarcastic humor and irony

D Prompts

This Appendix contains all the prompts used in the pipeline.

D.1 Initial prompt

System prompt:

```
##Instruction:
You will be given a document. Your task is to create a comprehensive
   list of descriptors - words or phrases that distill the meaning,
   tone, style, genre, topics, and other characteristics of the
   document. Do not focus solely on the topics; also include phrases
   that describe the tone and style.
The descriptors you generate should be useful for two purposes:
1. Rewriting the document in a way that is faithful to the original.
2. Classifying the document in a large variety of NLP classification
   tasks.
##Requirements:
1. Generate Descriptors:
- Describe the document in aspects including, but not limited to,
   style, tone, genre, topic, domain, language, quality, sentiment
   etc. Describe anything that you believe is essential for
   understanding, summarizing, or rewriting the document.
- If necessary, you can add minute details into the "specifics" list.
    This list can contain individual words, phrases and other details
    that do not necessarily describe the document but are important
   for rewriting it.
2. Descriptor Details:
- Descriptors can be single words or multi-word phrases. They should
   typically not be longer than 3 words.
- Each descriptor should describe only one aspect of the document.
   Multi-aspect descriptors should be split into separate descriptors
   .
- The list should be so extensive that, given only the list, a
   competent person should be able to rewrite the original document.
- Do not repeat descriptors and avoid near-duplicates.
3. Order:
- Order the descriptors in order of importance: the most important
   descriptors should be listed first.
- Importance is determined by how essential the descriptor is for
   reconstructing, understanding, or summarizing the document.
4. Descriptor Explanations:
- For each descriptor, explain its meaning and relevance to the
   document in 2 to 3 sentences.
- The items in the specifics list do not need to be explained.
5. Format:
- The output must be formatted as a JSON object. Follow this
   formatting exactly.
- Never add any preamble or anything else outside the JSON object.
- Separate the descriptors and their explanations with a semicolon
   ';'. Do NOT include a semicolon anywhere else in the descriptor.
- Example: {"descriptors": ["<descriptor>; <explanation>", "<
   descriptor>; <explanation>", "<descriptor>; <explanation>", ...],
```

```
   "specifics": ["<specifics>", "<specifics>", "<specifics>",...]}
```

User prompt:

```
    {document}
```

D.2 Rewrite prompt

System prompt:

```
##Instruction:
You will be given a list of descriptors and specifics:
- Descriptors: Describe the document in aspects including, but not
   limited to, style, tone, genre, topic, domain, language, quality,
   sentiment, etc.
- Each descriptor is followed by an explanation of its meaning and
   relevance to the document.
- The descriptors are ordered in order of importance: the most
   important descriptors are listed first.
- Specifics: Minute details such as individual words and phrases,
   emphasis, structure, etc. to help in rewriting the document.
##Your Task:
Reconstruct the original document as accurately as possible using
   both the descriptors and specifics provided. Do not describe the
   document; instead, produce the document itself.
##Requirements:
1. Utilize All Descriptors:
- Incorporate all descriptors and specifics into your writing.
- Ensure that the content, tone, style, and structure align with the
   descriptors.
- The descriptors are ordered in order of importance: the most
   important descriptors are listed first.
2. Format:
- The output must be formatted as a JSON object. Follow this
   formatting exactly.
- Never add any preamble or anything else outside the JSON object.
- Example: {{"text": "<rewritten version of the document>"}}
```

User prompt:

```
*Descriptors*:
<start>{descriptors}<end>
*Specifics*:
<start>{specifics}<end>
```

D.3 Revision prompt

System prompt:

```
##Instruction:
You will be given:
1. Two Documents:
- Original Document: The initial text.
- Rewritten Document: An attempt to recreate the original document
   based solely on provided descriptors and specifics.
```

```
2. A List of Descriptors and Specifics (describing the original
   document):
- Descriptors: Describe the document in aspects including, but not
   limited to, style, tone, genre, topic, domain, language, quality,
   sentiment, etc.
- Each descriptor is followed by an explanation of its meaning and
   relevance to the document.
- The descriptors are ordered in order of importance: the most
   important descriptors are listed first.
- Specifics: Minute details such as individual words and phrases,
   emphasis, structure, etc. to help in rewriting the document.
##Your Task:
1. Critique the Rewritten Document:
- Analyze the rewritten document.
- Point out ways in which the rewrite differs from the original.
   Point out specific sentences or phrases that diverge meaningfully
   from the original, if applicable.
- Focus on discrepancies in content, tone, style, structure, and any
   missing or misrepresented elements.
2. Revise the Descriptors and Specifics:
- Adjust and revise the descriptors and specifics to help in creating
    a more faithful rewrite of the original document and for
   classifying it in a large variety of NLP classification tasks.
- Each descriptor should describe only one aspect of the document.
   Multi-aspect descriptors should be split into separate descriptors
   .
- The specifics list can contain individual words, phrases and other
   details that do not necessarily describe the document but are
   important for rewriting it.
- The revised lists should not contain any direct or indirect
   references or comparisons to the original lists. They should stand
    on their own.
- For each descriptor, explain its meaning and relevance to the
   document in 2 to 3 sentences. The specifics do not need to be
   explained.
##Requirements:
1. Specificity: Provide detailed observations in your critique.
2. Clarity: Ensure the revised descriptors are clear and precise.
3. Completeness: Cover all aspects where the rewritten document
   deviates from the original.
4. Conciseness: Do not make the descriptors overly long. They should
   typically not be longer than 3 words. Do not repeat descriptors
   and avoid near-duplicates.
5. Order: Order the descriptors in order of importance: the most
   important descriptors should be listed first.
6. Format:
- The output must be formatted as a JSON object. Follow this
   formatting exactly.
- Never add any preamble or anything else outside the JSON object.
- Separate the descriptors and their explanations with a semicolon
   ';'. Do NOT include a semicolon anywhere else in the descriptor.
```

```
- Example: {"differences": "<differences between original and rewrite
   >",
"descriptors": ["<descriptor>; <explanation>", "<descriptor>; <
   explanation>", "<descriptor>; <explanation>", ...], "specifics":
   ["<specifics>", "<specifics>", "<specifics>", ...]}
```

User prompt:

```
{
*Original Document*:
<start>{original_document}<end>
*Rewritten Document*:
<start>{rewritten_document}<end>
*Descriptors*:
<start>{descriptors}<end>
*Specifics*:
<start>{specifics}<end>
}
```

D.4 Disambiguation prompt

System prompt:

```
You will be given a descriptor and its list of explainers. The
   descriptor describes a document and the explainers provide
   additional context to the descriptor.
Each explainer line is prefixed with its INDEX in square brackets,
   like: [0] text...
Your job is to merge/split explainers into semantically consistent
   groups *and return the INDICES of the original explainers you
   placed into each group*. Do not invent new indices, and use each
   index exactly once across all groups.
\#\# Instructions
1) Read the descriptor and explainers (each line is like: [<index>] <
   explainer text>).
2) Compare the explainers to each other and consider:
    - Do they point to the same kind of interpretation, rationale, or
        feature?
    - Do they reflect a consistent understanding or application of
       the descriptor?
3) Based on this comparison, create one or more semantically coherent
    groups.
4) For each group, provide:
   - group_descriptor (<= 3 words where possible. Coherent with the
      explainers in this group.)
   - group_explainer (Concise, canonical summary that covers the
      meaning of all explainers in this group)
   - original_explainer_indices_in_this_group: an array of the *
      indices* you grouped here (e.g., [0, 5, 7]).
Notes:
- Do NOT drop any original explainer indices. Every original
   explainer index must appear exactly once across groups.
```

```
- You may keep the original descriptor for one group; newly invented
   descriptors are allowed for other groups.
- Make minimal changes to the original explainers when forming
   group_explainers; focus on summarizing and unifying the meanings.
- Descriptors and explainers must be self-contained -- they must not
   contain references or comparisons to the original descriptors or
   explainers or the other groups.
\#\# Output format
Return a single valid JSON object with this exact structure (no extra
    text or commentary of any kind):
{"original_descriptor": "<original descriptor>",
  "groups": [
    {"group_descriptor": "<group descriptor>", "group_explainer": "<
       group explainer>", "original_explainer_indices_in_this_group":
        [<index_a>, <index_b>, ...]}
  ]}
```

User prompt:

```
Descriptor: {descriptor}
Explainers (each prefixed by its INDEX):
{explainers}
```

D.5 Synonym merging prompt

System prompt:

```
You are deciding whether two descriptor-explainer pairs, A and B,
   refer to the same concept.
The descriptors describe a document, while the explainer provides
   additional context.

Treat them as synonyms (set is_synonym=true) if any of the following
   is true:
    - They refer to the same concept or entity despite different
       wording.
    - They deviate slightly in specificity but still cover the same
       core meaning.
    - They differ in phrasing but could be used interchangeably in
       most contexts.
    - They differ in word order, punctuation or spelling but mean the
        same thing.
    - They list similar things in different order.

Do NOT treat them as synonyms (set is_synonym=false) if any of the
   following is true:
    - They refer to different concepts/entities or mutually exclusive
        categories.
    - One is substantially broader/narrower in scope or specificity.
    - The explainer introduces a difference in meaning even when
       descriptors are similar.

In edge cases, prefer to NOT treat as synonyms.
```

```
If synonyms:
    - choose which to keep: "A" or "B" (prefer clearer, more general
       phrasing);
    - set keep to the chosen label and drop to the other.

If not synonyms:
    - Still return keep and drop as A and B (order doesn't matter).

Return ONLY JSON with keys: is_synonym (bool), keep ("A"|"B"), drop
   ("A"|"B").
Do not rewrite or invent descriptors.
```

User prompt:

```
A: descriptor: {descriptorA} explainer: {explainerA}
==============
B: descriptor: {descriptorB} explainer: {explainerB}
```

D.6 Force merging prompt

System prompt:

```
You are selecting a representative pair among candidates that share
   the same descriptor.
The descriptor describes a document, while the explainer provides
   additional context.
Your task is to choose the most representative pair among the given
   pairs.
Return the integer index of the pair to keep in the field 'keep'
When choosing the representative pair, prefer concise, general,
   commonly used phrasing that covers the meaning of all pairs.
Return ONLY JSON with: keep (int). Do not invent or rewrite anything.
```

User prompt:

```
index: {i} | descriptor: {descriptorA} | explainer: {explainerA}
index: {j} | descriptor: {descriptorA} | explainer: {explainerB}
...
```

D.7 Harmonization prompt

System prompt:

```
You are a careful ontology assistant.
Task: Given a query pair in the form 'descriptor; explainer' and a
   list of schema candidates, pick the single candidate that is
   synonymous or near-synonymous in most contexts to the query.
If many candidates are synonymous, choose the one that is the closest
    match.
If you find a suitable candidate, respond with is_synonym=true and
   set chosen_index to exactly the chosen candidate's index.
If none are acceptable, return is_synonym=false.
Output MUST be JSON only.
```

User prompt:

```
Query: {descriptor}; {explainer}
Candidates:
{i} {candidateA}
{j} {candidateB}
...
```

E Annotation results

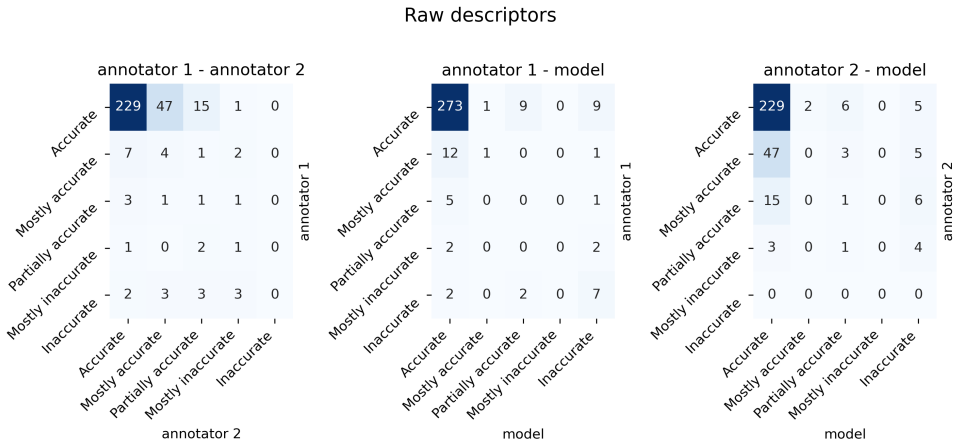


Figure 6: Confusion matrices for descriptor accuracy annotations with raw descriptors.

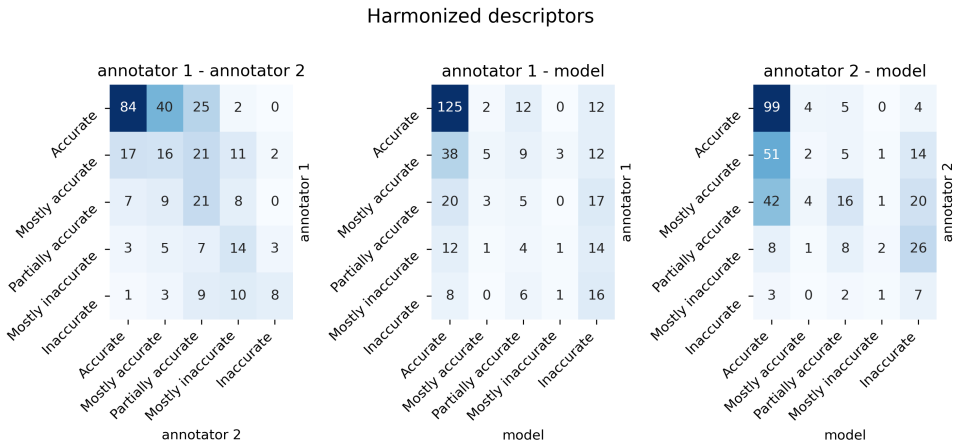


Figure 7: Confusion matrices for descriptor accuracy annotations with harmonized descriptors.

F Descriptor aspect categories

Main category	Examples
Content & Subject Matter	e.g. topic, domain, subtopic, intent, cultural context, temporal relevance
Structure & Format	e.g. genre, document type (blog, news, etc.), media format, length
Style & Tone	e.g. formality, tone, technicality (e.g. jargon), voice (active/passive), syntax
Quality & Trustworthiness	e.g. factuality, authority, originality, spam, bias, depth (superficial/in-depth)
Audience & Engagement	e.g. target audience, engagement level, monetization, call-to-action, shareability
Linguistic Features	e.g. readability, sentiment, figurative language, multilinguality, repetition
Ethical & Legal	e.g. toxicity, misinformation, legal compliance, accessibility, dark patterns
Other	miscellaneous, not fitting into other categories

Table 5: Broad categories in which we categorize descriptors with examples of their contents

The descriptor categories were created through an iterative process involving discussion between the authors and with the help of an LLM. The categories reflect the type of descriptors that we observed to exist in the data and the kinds of descriptors that we would hope to exist and that researchers in the field of corpus linguistics and NLP might find relevant.

G Classification results

Task	Features	Macro-F1 [95% CI]	W-F1	$\|V\|$
FineWeb Edu	Descriptors (Harmonized)	0.70 [0.69, 0.71]	0.92	278,093
FineWeb Edu	Descriptors (Raw)	0.76 [0.75, 0.77]	0.93	351,003
FineWeb Edu	TF-IDF	0.75 [0.74, 0.77]	0.94	50,000
Topic	Descriptors (Harmonized)	0.63 [0.62, 0.64]	0.64	278,093
Topic	Descriptors (Raw)	0.65 [0.64, 0.65]	0.66	351,003
Topic	TF-IDF	0.73 [0.72, 0.74]	0.75	50,000
Format	Descriptors (Harmonized)	0.44 [0.43, 0.45]	0.53	278,093
Format	Descriptors (Raw)	0.48 [0.47, 0.49]	0.57	351,003
Format	TF-IDF	0.49 [0.48, 0.50]	0.61	50,000

Table 6: Classification F1 comparing harmonized vs. raw descriptors, with TF-IDF as a text-based reference. Confidence intervals from 1000 bootstrap resamples of the test set. Descriptor configs use minimum frequency 1 (full vocabulary); TF-IDF capped at 50,000 features. $|V|$ = vocabulary size.

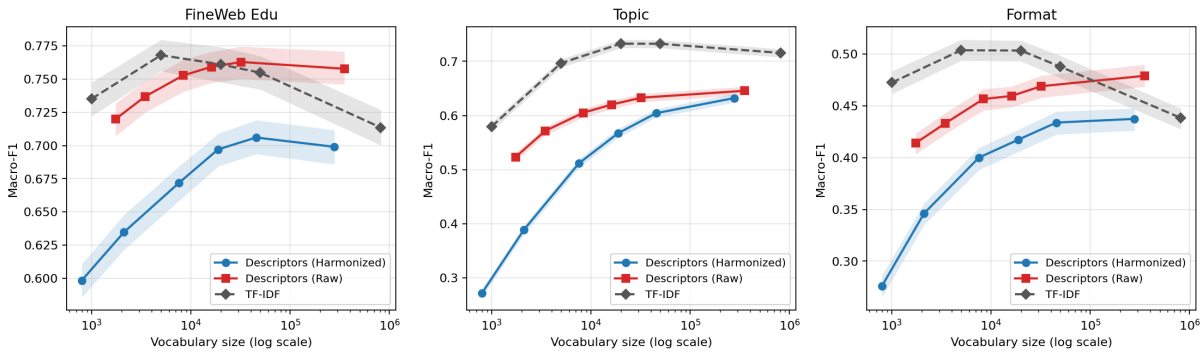


Figure 8: Macro-F1 as a function of feature-set size for harmonized descriptors, raw descriptors, and TF-IDF, across the three tasks. Descriptor vocabulary size is swept by minimum token frequency; TF-IDF by maximum vocabulary size. Shaded bands are 95% bootstrap confidence intervals.

Class	Desc. (Harm.) Scan%	Lift	Prec.	Desc. (Raw) Scan%	Lift	Prec.	TF-IDF Scan%	Lift	Prec.	N
edu	16.9%	4.7	37.9%	14.5%	5.5	44.0%	10.9%	7.4	58.7%	1,597
not-edu	74.4%	1.1	98.9%	74.3%	1.1	99.1%	73.8%	1.1	99.7%	18,403
Weighted mean	69.8%	1.4	94.0%	69.5%	1.4	94.7%	68.8%	1.6	96.5%	20,000

Table 7: Per-class retrieval at 80% recall, FineWeb Edu. Scan% = fraction of test set scanned to reach the target recall (lower better); Lift = precision/base rate; Prec. = precision at that point; N = support. Min. frequency 1; TF-IDF capped at 50,000. Sorted by harmonized Scan%.

Class	Desc. (Harm.) Scan%	Lift	Prec.	Desc. (Raw) Scan%	Lift	Prec.	TF-IDF Scan%	Lift	Prec.	N
edu	2.0%	9.8	77.9%	2.0%	10.2	81.8%	1.7%	11.9	94.7%	1,597
not-edu	18.4%	1.1	99.9%	18.4%	1.1	100.0%	18.4%	1.1	100.0%	18,403
Weighted mean	17.1%	1.8	98.1%	17.1%	1.8	98.5%	17.1%	1.9	99.5%	20,000

Table 8: Per-class retrieval at 20% recall, FineWeb Edu. Columns as in Table 7.

Class	Desc. (Harm.) Scan%	Lift	Prec.	Desc. (Raw) Scan%	Lift	Prec.	TF-IDF Scan%	Lift	Prec.	N
Adult	2.2%	35.7	26.1%	5.6%	14.2	10.4%	2.5%	32.3	23.6%	146
Religion	4.2%	18.9	39.2%	3.8%	21.3	44.0%	3.0%	26.3	54.4%	414
Travel	4.5%	17.8	42.3%	4.3%	18.5	43.9%	3.0%	26.7	63.2%	474
Software Dev.	4.5%	17.6	29.6%	4.9%	16.5	27.7%	2.9%	27.4	46.1%	336
History	7.4%	10.8	13.0%	4.6%	17.5	21.1%	3.4%	23.4	28.2%	241
Crime & Law	8.0%	10.0	35.9%	7.7%	10.4	37.2%	4.9%	16.4	59.1%	719
Hardware	8.3%	9.7	24.6%	8.1%	10.0	25.3%	3.9%	20.5	52.0%	508
Games	8.6%	9.3	28.2%	10.0%	8.0	24.3%	5.8%	13.9	42.4%	608
Fashion & Beauty	8.7%	9.2	34.9%	9.2%	8.7	33.0%	4.2%	18.9	71.9%	760
Science & Tech.	8.8%	9.1	36.9%	9.0%	8.9	36.1%	6.7%	12.0	48.4%	808
Food & Dining	9.2%	8.7	38.7%	10.0%	8.0	35.8%	4.5%	18.0	80.3%	894
Health	9.7%	8.3	50.7%	8.5%	9.4	57.4%	6.7%	11.9	73.0%	1,225
Software	11.0%	7.3	21.0%	9.5%	8.4	24.3%	5.8%	13.8	39.9%	579
Sports & Fitness	11.2%	7.1	56.8%	10.6%	7.5	59.9%	7.7%	10.4	83.1%	1,595
Politics	11.3%	7.1	46.2%	11.1%	7.2	47.1%	8.5%	9.4	61.3%	1,305
Entertainment	11.8%	6.8	50.6%	14.0%	5.7	42.7%	8.4%	9.5	70.9%	1,497
Education & Jobs	12.8%	6.3	37.0%	12.9%	6.2	36.7%	7.6%	10.6	62.3%	1,180
Industrial	13.3%	6.0	17.5%	11.3%	7.1	20.6%	6.1%	13.1	37.7%	578
Literature	13.7%	5.8	22.4%	14.2%	5.6	21.6%	8.3%	9.6	36.8%	766
Art & Design	14.4%	5.5	17.2%	10.5%	7.6	23.6%	5.0%	15.9	49.4%	620
Finance & Business	16.0%	5.0	37.5%	14.7%	5.5	41.0%	10.4%	7.7	57.7%	1,503
Home & Hobbies	16.7%	4.8	32.9%	17.4%	4.6	31.6%	10.0%	8.0	54.9%	1,374
Transportation	16.8%	4.8	16.0%	18.2%	4.4	14.8%	5.0%	15.8	53.2%	672
Social Life	18.7%	4.3	25.7%	17.0%	4.7	28.2%	12.7%	6.3	37.8%	1,198
Weighted mean	11.8%	7.8	36.3%	11.6%	7.9	36.8%	7.2%	12.9	58.4%	20,000

Table 9: Per-class retrieval at 80% recall, Topic. Columns as in Table 7.

Class	Desc. (Harm.)			Desc. (Raw)			TF-IDF			N
	Scan%	Lift	Prec.	Scan%	Lift	Prec.	Scan%	Lift	Prec.	
Adult	0.1%	132.6	96.8%	0.2%	128.4	93.8%	0.2%	128.4	93.8%	146
History	0.3%	63.5	76.6%	0.3%	58.1	70.0%	0.3%	65.6	79.0%	241
Software Dev.	0.5%	44.0	73.9%	0.5%	40.9	68.7%	0.4%	46.0	77.3%	336
Religion	0.5%	43.1	89.2%	0.5%	39.7	82.2%	0.5%	42.7	88.3%	414
Travel	0.6%	36.8	87.2%	0.6%	33.4	79.2%	0.5%	39.3	93.1%	474
Hardware	0.6%	33.7	85.7%	0.7%	28.9	73.4%	0.6%	32.6	82.9%	508
Games	0.6%	31.4	95.3%	0.7%	29.5	89.7%	0.6%	32.4	98.4%	608
Art & Design	0.8%	25.8	80.0%	0.9%	22.9	70.9%	0.7%	28.6	88.6%	620
Crime & Law	0.8%	25.2	90.6%	0.8%	24.3	87.3%	0.8%	26.5	95.4%	719
Fashion & Beauty	0.8%	24.1	91.6%	0.8%	23.5	89.4%	0.8%	26.0	98.7%	760
Transportation	0.9%	22.8	76.7%	1.0%	20.0	67.2%	0.8%	26.1	87.7%	672
Food & Dining	0.9%	21.8	97.3%	0.9%	21.8	97.3%	0.9%	21.9	97.8%	894
Science & Tech.	0.9%	21.6	87.1%	1.1%	18.7	75.3%	0.9%	21.4	86.6%	808
Software	1.0%	20.9	60.4%	1.0%	19.8	57.4%	0.8%	24.6	71.2%	579
Industrial	1.0%	20.6	59.5%	1.0%	19.5	56.3%	0.7%	26.9	77.8%	578
Literature	1.3%	15.3	58.8%	1.3%	15.5	59.5%	0.9%	22.8	87.5%	766
Health	1.4%	14.7	90.1%	1.4%	14.7	90.1%	1.3%	15.3	93.5%	1,225
Education & Jobs	1.4%	14.5	85.8%	1.4%	14.1	83.1%	1.3%	15.5	91.5%	1,180
Politics	1.6%	12.7	82.9%	1.6%	12.1	78.8%	1.4%	14.2	92.9%	1,305
Sports & Fitness	1.6%	12.5	100.0%	1.6%	12.3	98.5%	1.6%	12.4	99.1%	1,595
Entertainment	1.6%	12.3	92.3%	1.7%	12.0	89.8%	1.6%	12.9	96.8%	1,497
Home & Hobbies	1.7%	11.8	81.4%	1.8%	11.3	77.7%	1.5%	13.6	93.5%	1,374
Finance & Business	1.7%	11.6	87.0%	1.9%	10.9	81.6%	1.6%	12.4	93.5%	1,503
Social Life	2.0%	9.8	58.8%	1.9%	10.4	62.5%	1.4%	13.9	83.0%	1,198
Weighted mean	1.3%	19.8	83.8%	1.3%	18.8	80.1%	1.1%	21.5	91.2%	20,000

Table 10: Per-class retrieval at 20% recall, Topic. Columns as in Table 7.

Class	Desc. (Harm.)			Desc. (Raw)			TF-IDF			N
	Scan%	Lift	Prec.	Scan%	Lift	Prec.	Scan%	Lift	Prec.	
Legal Notices	1.5%	55.4	34.4%	1.2%	69.2	42.9%	1.2%	65.0	40.3%	124
Documentation	4.6%	17.5	8.9%	3.5%	22.8	11.6%	1.4%	58.0	29.6%	102
Academic Writing	5.0%	16.2	25.9%	3.5%	22.6	36.1%	2.9%	27.6	44.1%	319
Creative Writing	5.9%	13.6	16.4%	5.3%	15.2	18.4%	7.1%	11.2	13.6%	242
User Review	9.8%	8.1	15.2%	7.9%	10.1	18.8%	7.7%	10.4	19.3%	372
Comment Section	12.3%	6.5	24.5%	12.1%	6.6	24.9%	8.1%	9.9	37.0%	751
About (Pers.)	13.0%	6.2	11.4%	10.1%	7.9	14.6%	5.3%	15.2	28.1%	369
Tutorial	14.6%	5.5	25.2%	11.0%	7.3	33.4%	8.3%	9.6	44.3%	920
Customer Support	15.7%	5.1	8.1%	12.8%	6.3	9.9%	7.2%	11.2	17.8%	318
About (Org.)	16.1%	5.0	17.5%	14.9%	5.4	18.9%	9.5%	8.4	29.7%	704
Spam / Ads	20.5%	3.9	9.3%	14.9%	5.4	12.7%	4.9%	16.4	38.8%	473
Q&A Forum	22.5%	3.6	2.7%	14.0%	5.8	4.4%	6.0%	13.4	10.2%	153
News Article	23.2%	3.5	53.3%	22.0%	3.6	56.1%	16.7%	4.8	74.3%	3,092
News (Org.)	23.7%	3.4	20.8%	20.6%	3.9	23.8%	16.5%	4.9	29.8%	1,227
Personal Blog	23.7%	3.4	50.2%	25.9%	3.1	45.9%	24.0%	3.3	49.5%	2,974
Knowledge Article	24.4%	3.3	14.4%	20.5%	3.9	17.1%	11.8%	6.8	29.6%	877
Product Page	25.1%	3.2	54.9%	26.1%	3.1	52.9%	21.4%	3.7	64.4%	3,444
Structured Data	25.3%	3.2	5.1%	24.6%	3.2	5.2%	9.5%	8.4	13.4%	320
Audio Transcript	28.1%	2.8	4.3%	23.6%	3.4	5.1%	13.4%	6.0	8.9%	300
Nonfiction Writing	31.0%	2.6	14.9%	29.5%	2.7	15.6%	19.7%	4.1	23.4%	1,152
Listicle	34.7%	2.3	5.7%	25.9%	3.1	7.6%	19.0%	4.2	10.4%	492
Truncated	47.6%	1.7	3.1%	41.7%	1.9	3.6%	13.5%	5.9	11.0%	371
FAQ	49.2%	1.6	0.7%	29.9%	2.7	1.1%	13.4%	6.1	2.5%	84
Content Listing	52.5%	1.5	6.2%	44.4%	1.8	7.4%	24.4%	3.3	13.4%	820
Weighted mean	23.9%	4.3	33.0%	22.3%	4.9	33.8%	16.3%	6.8	43.7%	20,000

Table 11: Per-class retrieval at 80% recall, Format. Columns as in Table 7.

Class	Desc. (Harm.) Scan%	Lift	Prec.	Desc. (Raw) Scan%	Lift	Prec.	TF-IDF Scan%	Lift	Prec.	N
Legal Notices	0.1%	161.3	100.0%	0.1%	155.1	96.2%	0.1%	155.1	96.2%	124
Documentation	0.3%	79.2	40.4%	0.1%	152.5	77.8%	0.2%	117.6	60.0%	102
FAQ	0.3%	68.6	28.8%	1.2%	17.5	7.4%	0.2%	88.0	37.0%	84
Audio Transcript	0.4%	51.9	77.9%	0.6%	34.5	51.7%	0.4%	53.3	80.0%	300
Creative Writing	0.4%	47.6	57.6%	0.5%	43.5	52.7%	0.3%	75.0	90.7%	242
Academic Writing	0.5%	45.1	71.9%	0.4%	50.8	81.0%	0.4%	51.4	82.0%	319
Q&A Forum	0.6%	36.5	27.9%	0.7%	27.2	20.8%	0.3%	63.3	48.4%	153
User Review	0.6%	31.8	59.1%	0.7%	28.2	52.4%	0.6%	32.5	60.5%	372
About (Pers.)	0.7%	31.1	57.4%	0.6%	33.1	61.2%	0.5%	37.8	69.8%	369
Spam / Ads	0.7%	29.3	69.3%	0.8%	24.8	58.6%	0.5%	37.9	89.6%	473
Customer Support	0.8%	24.8	39.5%	0.8%	25.0	39.8%	0.5%	38.3	61.0%	318
Structured Data	0.8%	24.4	39.0%	0.8%	23.4	37.4%	0.4%	48.2	77.1%	320
Comment Section	1.0%	19.8	74.4%	1.0%	19.9	74.8%	0.8%	23.7	88.8%	751
Tutorial	1.1%	18.0	82.9%	1.1%	17.8	81.8%	1.1%	18.3	84.4%	920
About (Org.)	1.4%	14.8	52.0%	1.4%	14.3	50.4%	1.0%	19.4	68.5%	704
Knowledge Article	1.6%	12.4	54.3%	1.7%	11.9	52.4%	1.2%	16.4	72.1%	877
News (Org.)	1.8%	10.9	66.8%	2.0%	9.8	60.0%	1.7%	11.9	73.0%	1,227
Listicle	2.0%	10.0	24.5%	2.1%	9.6	23.7%	1.2%	16.2	39.8%	492
Content Listing	2.7%	7.5	30.8%	3.4%	5.8	23.8%	1.8%	11.1	45.4%	820
Truncated	3.0%	6.8	12.7%	2.5%	8.0	14.8%	0.5%	37.4	69.4%	371
Nonfiction Writing	3.1%	6.6	37.8%	2.9%	6.9	39.6%	2.1%	9.7	55.8%	1,152
News Article	3.6%	5.6	86.6%	3.8%	5.3	82.5%	3.2%	6.2	95.5%	3,092
Personal Blog	3.6%	5.5	82.3%	3.7%	5.4	80.5%	3.4%	5.9	88.1%	2,974
Product Page	3.8%	5.2	89.8%	4.0%	5.1	87.0%	3.8%	5.3	91.9%	3,444
Weighted mean	2.6%	13.6	69.5%	2.7%	13.1	66.8%	2.2%	17.3	80.1%	20,000

Table 12: Per-class retrieval at 20% recall, Format. Columns as in Table 7.

H Full query results

Format	Harmonized Prec.	Harmonized Rec.	Raw Prec.	Raw Rec.	Topic	Harmonized Prec.	Harmonized Rec.	Raw Prec.	Raw Rec.
Tutorial	0.42	0.47	0.51	0.45	Finance & Business	0.58	0.45	0.67	0.36
Creative Writing	0.08	0.35	0.11	0.19	Home & Hobbies	0.69	0.23	0.72	0.20
News (Org.)	0.37	0.29	0.46	0.17	Transportation	0.60	0.38	0.65	0.19
About (Org.)	0.23	0.02	0.33	0.04	Politics	0.65	0.52	0.74	0.42
Structured Data	0.36	0.04	0.34	0.07	History	0.19	0.50	0.28	0.38
User Review	0.26	0.13	0.39	0.09	Literature	0.59	0.20	0.73	0.13
Personal Blog	0.46	0.68	0.57	0.61	Games	0.68	0.40	0.80	0.38
Content Listing	0.40	0.04	0.47	0.05	Art & Design	0.40	0.18	0.45	0.15
Knowledge Article	0.10	0.52	0.12	0.39	Science & Tech.	0.43	0.47	0.54	0.37
News Article	0.52	0.24	0.70	0.22	Crime & Law	0.64	0.45	0.76	0.41
Truncated	0.52	0.02	0.56	0.02	Health	0.52	0.68	0.60	0.62
FAQ	0.31	0.23	0.28	0.31	Entertainment	0.67	0.32	0.71	0.28
About (Pers.)	0.07	0.34	0.11	0.40	Education & Jobs	0.64	0.34	0.72	0.26
Product Page	0.60	0.27	0.64	0.28	Industrial	0.44	0.39	0.49	0.31
Spam / Ads	0.13	0.04	0.17	0.07	Fashion & Beauty	0.78	0.44	0.85	0.28
Nonfiction Writing	0.12	0.21	0.09	0.11	Software Dev.	0.53	0.37	0.63	0.30
Academic Writing	0.13	0.58	0.37	0.47	Travel	0.54	0.60	0.68	0.38
Q&A Forum	0.22	0.18	0.28	0.15	Adult	0.60	0.57	0.67	0.66
Comment Section	0.51	0.20	0.67	0.19	Social Life	0.39	0.23	0.39	0.21
Audio Transcript	0.41	0.32	0.47	0.25	Food & Dining	0.80	0.61	0.82	0.52
Documentation	0.07	0.77	0.09	0.67	Software	0.34	0.24	0.38	0.19
Legal Notices	0.30	0.53	0.33	0.46	Religion	0.48	0.71	0.52	0.62
Customer Support	0.07	0.32	0.14	0.24	Sports & Fitness	0.78	0.37	0.84	0.32
Listicle	0.54	0.06	0.57	0.05	Hardware	0.63	0.33	0.70	0.34

Class	Desc. (Harm.)	Desc. (Raw)	TF-IDF
edu	general context (35), development narratives (18), scientific-research (308), informing the reader (37), use of etymology (19), word history (26), holiday history (8), anti-bullying strategy (9), discovery and artifact context (7), teacher reflection (11)	scientific topic (53), educational resource (94), educational topic (45), historical account (212), scientific language (279), environmental topic (59), educational content (695), historical event (84), historical narrative (119), scientific study (143)	or (46,836), of the (53,488), species (1,197), century (2,786), of (75,396), scientists (1,195), researchers (1,416), nasa (345), used (14,612), children (6,776)
not-edu	product description (1,028), international travel guide (237), customer centricity (504), sports article (489), local news update (239), opinionated writing (196), local news (612), professional language use (762), tone - promotional (2,231), personal story (241)	home improvement (284), sports news (1,681), product description (2,078), varied topics (177), travel guide (473), financial (232), professional language (1,292), business news (493), respectful dialogue (92), customer-centric (1,279)	my (21,750), me (16,737), 2018 (2,622), 2014 (3,109), our (29,795), we (38,363), business (10,558), 2019 (2,062), am (10,094), 2016 (3,171)

Table 13: Top-10 features per class for FineWeb Edu, ranked by logistic regression coefficient, comparing harmonized descriptors, raw descriptors, and TF-IDF. Parenthesised numbers are training-set document frequencies. Descriptor tags shown; definitions omitted. Descriptors at min. frequency 1; TF-IDF capped at 50,000.

PAWN++ in the Wild: Evaluating AI-Generated Text Detection on the Web

Oleg Chernikov **Evgenii Ptitsyn** **Albina Burlova**
HSE University
{ovchernikov,egptitsyn}@edu.hse.ru
aburlova@hse.ru

Abstract

Machine-generated text is becoming a routine part of the web. Most detectors, however, are still evaluated on curated datasets rather than real web content. We introduce PAWN++, a compact supervised detector that learns which likelihood signals from two frozen language models are most informative. On MAGE, our method reaches 98.36 AUROC in distribution and 98.42 under a joint domain and generator shift. It also transfers well to external benchmarks without target-specific training.

To test whether this robustness extends to the web, we construct WATTLE, a benchmark of 41,300 texts in which each machine output is tied to its original human web page. The resulting data capture realistic transformations of web content as well as variation across languages and time. We use WATTLE to compare our method with seven other detectors. Performance deteriorates sharply on realistic rewrites: our method reaches 70.1 AUROC overall and 58.1 on text produced by a frontier model. Likelihood-based detectors fall below chance, and no evaluated system identifies more than 12% of frontier rewrites at a 1% false-positive rate. These results expose a substantial gap between standard evaluation scores and reliable detection on the web.[1]

1 Introduction

Large language models write fluent news, reviews, answers, and essays. Their output now flows into the same channels as human writing, and the ability to tell the two apart underpins moderation, provenance, and the curation of training corpora. Detection research has produced strong supervised classifiers and elegant zero shot statistics, and it evaluates them almost exclusively on curated benchmarks assembled from clean sources such as news wires, Wikipedia, and question answering forums (Li et al., 2024; Dugan et al., 2024; Artemova et al., 2025).

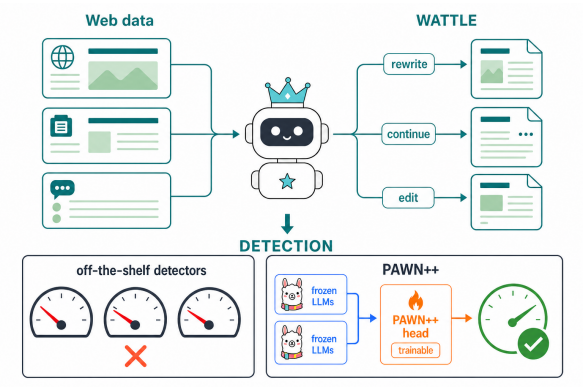


Figure 1: WATTLE anchors machine rewrites, continuations, and edits to real web pages. Likelihood ratio detectors fall below chance on frontier text, and the learned head of PAWN++ reads two frozen probes to stay above chance where fixed statistics invert.

This paper follows one detector along the full path from a curated benchmark to the open web. We first present PAWN++, an extension of the Perplexity Attention Weighted Network (Miralles-González et al., 2026) that reads token level likelihood evidence from two related frozen language models through a learned head of 2.7M trainable parameters, and it delivers the highest average recall in its MAGE comparison (Li et al., 2024) with competitive in distribution AUROC, and the best transfer to unseen domains and generators. Applied without any adaptation, the same checkpoint leads the no training comparison on RAID (Dugan et al., 2024) at a fixed false positive budget, attains the strongest AUROC in its HC3 comparison in both English and Chinese (Guo et al., 2023), a language it never saw in training, and stays competitive on Beemo (Artemova et al., 2025).

The last and hardest step is the web itself. Web pages are noisy, multilingual, and unlabeled, and

[1]All data and code are available at https://github.com/OlegHSE/PAWN-in-the-Wild

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 282–296
October 29, 2026 ©2026 Association for Computational Linguistics

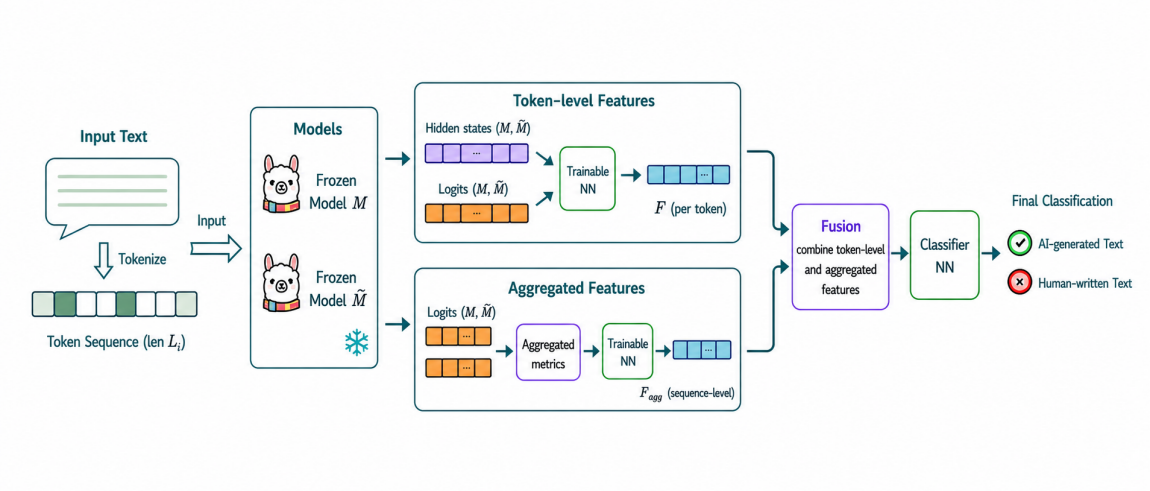


Figure 2: PAWN++ architecture. Two frozen LLMs contribute token probabilities and hidden states; learned PAWN modules combine token, cross-model, and sequence-level evidence.

machine text on the web comes from frontier commercial models. We construct WATTLE, a parallel benchmark of 41,300 objects in which every machine text is anchored to a real web page whose human version predates the modern generation era, covering rewriting, continuation, editing, six languages, and two temporal crawl slices. The results reorder the field. Frozen likelihood ratio and curvature detectors fall below chance on frontier rewrites, supervised detectors keep useful rankings yet collapse at deployment thresholds, and calibrated false positive budgets fail to survive language transfer for every frozen method. A final experiment trains the head anew on five thousand web texts from the benchmark pipeline and shows how misleading in protocol training can be. Scores jump to 96.4 AUROC on rewriting, and the same head trails the frozen checkpoint by 10 to 26 AUROC points on external machine text, so the gains reflect the generation protocol rather than detection.

Our contributions are:

- **PAWN++**, a compact supervised detector that learns to combine token- and sequence-level evidence from two frozen language models. It reaches 98.36 AUROC on MAGE and 98.42 under a joint domain and generator shift, while transferring well to external benchmarks.
- **WATTLE**, a benchmark of 41,300 texts in which every machine output is tied to its original human web page. It captures realistic transformations together with variation across languages and time.
- **A controlled web evaluation** of our method and seven representative baselines. Performance drops sharply on frontier-generated rewrites, where likelihood-based detectors fall below chance and no evaluated method recalls more than 12% at a 1% false-positive rate.

2 Related Work

Supervised detection fine tunes a pretrained encoder into a binary classifier (Liu et al., 2019; Tufts et al., 2025). These models reach high in domain accuracy and lose it under domain, generator, and paraphrase shift (Murgia et al., 2026; Pu et al., 2023; Akhilesh et al., 2025), and RADAR hardens the classifier adversarially against paraphrasing (Hu et al., 2023).

Zero shot detection reads statistics of model likelihoods. DetectGPT uses the curvature of the log probability surface (Mitchell et al., 2023), Fast-DetectGPT replaces its perturbations with an efficient sampling discrepancy (Bao et al., 2024), and Binoculars scores a text by the ratio of perplexity to cross perplexity between two related models (Hans et al., 2024). Follow up work adds token level hidden state discrepancy (Zhu et al., 2026) and temporal or spectral statistics of the surprisal sequence (Basani and Chen, 2026; Luo et al., 2026). The Perplexity Attention Weighted Network sits between the two families, learning how much each token contributes to likelihood features of one frozen model (Miralles-González et al., 2026). PAWN++ extends it with a second frozen probe, cross model metrics, uniform hidden state fusion, and sequence level conditioning (Perez et al., 2018), uniting the

cross model contrast of Binoculars with the surprisal dynamics view of DivEye inside one learned head.

Evaluation resources shaped this field. MAGE spans diverse sources and generators with out of distribution testbeds (Li et al., 2024), RAID stress tests detectors under adversarial attacks at fixed false positive budgets (Dugan et al., 2024), HC3 pairs human and ChatGPT answers in two languages (Guo et al., 2023), and Beemo studies machine text refined by experts and by stronger models (Artemova et al., 2025). A large scale audit of deployed detectors on social media documents reports unstable false positive behavior in the wild (Dutta et al., 2026). Closer to the web itself, WETBench measures task specific machine text on Wikipedia across three languages (Quaremba et al., 2025), DeGenTWeb identifies LLM dominant websites in Common Crawl and search results (He et al., 2026), temporal studies estimate the growth of machine content on the live internet (Dolezal et al., 2026), and detector based filtering of training corpora prevents language model collapse under recursive training (Drayson et al., 2025). WATTLE differs from these resources by anchoring every machine text to a real pre LLM web page, which yields source matched pairs, verified human gold, and an evaluation that reflects the web distribution rather than curated prose.

3 Method

3.1 Original PAWN

PAWN uses next-token distributions from a pretrained language model, but does not average their statistics uniformly. It learns how much each token should contribute to each likelihood-derived feature (Miralles-González et al., 2026).

LLM pass. A frozen decoder-only LLM processes input $x_{1:L}$ once, returning final hidden states and next-token logits. These outputs can be cached; only the lightweight PAWN networks are optimized.

Metric features. From the next-token distribution p_t, PAWN extracts entropy, maximum log-probability, observed-token log-probability, normalized rank, and cumulative top-p mass:

$$m_t = [\mathcal{H}_t, \max \log p_t, \log p_t(x_{t+1}), r_t, q_t], \quad (1)$$

Here $\mathcal{H}_t$ captures uncertainty, r_t is the observed token's rank, and q_t is the smallest top-p threshold that retains it. A metrics MLP maps m_t to z_t.

Configuration	AUROC
PAWN, Instruct	0.8753
PAWN, Base	0.8953
Two models + metrics	0.9126
Two models + metrics + XPPL	0.9096
Two models + metrics + XPPL + HS	0.9045
Two models + metrics + XPPL + uniform HS	0.9288
PAWN++	**0.9309**

Table 1: Ablations trained on 5,000 balanced examples and evaluated on the full MAGE test set.

Token weighting. At position t, a second MLP uses the current and next token hidden states plus relative position to produce the feature-wise importance-score vector a_t. A softmax over the sequence converts a_t into normalized relative weights, assigning each token a relative importance. The weights are then applied to metric features.

Aggregation and decision. PAWN pools the weighted metric features,

$$F = \sum_{t=1}^{L-1} \mathrm{softmax}_t(a_t) \odot z_t, \quad (2)$$

and a final MLP maps F to a binary logit. PAWN++ retains this metric–weight–aggregate pathway, then adds a second frozen model and sequence-level conditioning.

3.2 Second-Model and Cross-Model Features

Choice of the model pair. Following Binoculars (Hans et al., 2024), PAWN++ uses the Base and Instruct variants of the same model to capture complementary views of a given text. Their shared architecture and tokenizer make token-level likelihood and hidden-state comparisons directly comparable, so score differences reflect the text rather than model-family differences. The pair also provides a cross-model disagreement signal that a single model cannot produce. We do not claim that same-family pairs are optimal, heterogeneous pairs remain for future study.

PAWN++ runs `Llama-3.2-1B-Instruct` and `Llama-3.2-1B` in parallel with shared tokenization and frozen weights (Grattafiori et al., 2024). The same five token metrics are computed for each model and concatenated. We also include a cross-perplexity-style feature: the expectation of one model's negative log-probabilities under the other model's predictive distribution. Unlike standalone Binoculars, this cross-score is a token feature consumed by the learned PAWN metrics network.

Method	AUROC	Human R.	AI R.	Avg. R.
FastText	83.00	86.34	71.26	78.80
GLTR	74.00	12.42	**98.42**	55.42
Longformer (fine-tuned)	**99.00**	82.80	98.27	90.53
DetectGPT*	57.00	86.92	34.05	60.48
RADAR*‡	61.00	52.08	68.56	60.32
RADAR-FT	98.00	85.36	97.26	91.31
Binoculars*‡	65.00	10.60	8.23	9.42
LogRank	68.00	84.91	44.47	64.69
Entropy	67.00	75.68	50.04	62.86
DetectLLM	75.00	64.74	69.02	66.88
FastDetectGPT	70.00	93.65	41.73	67.69
BiScope	86.00	91.54	58.70	75.12
DivEye	88.00	73.72	82.57	78.15
PAWN (GPT2)	98.00	93.75	91.97	92.86
PAWN (Llama3-1B)	**99.00**	90.68	95.84	93.26
PAWN++ (ours)	98.36	**97.20**	93.11	**95.16**

Table 2: MAGE Testbed 4 (all source models and domains). Original baselines are from Miralles-González et al. (2026); additional methods are from Basani and Chen (2026); PAWN++ is our result. * unsupervised; ‡ threshold not selected.

This design represents cross-model disagreement together with each model's confidence structure rather than treating cross-perplexity as a sufficient statistic.

3.3 Uniform Hidden-State Fusion

The original PAWN weights network uses the last hidden layer. We instead normalize and average all K selected layers for model j:

$$\widetilde{h}_t^{(j)} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{LayerNorm}\left(h_t^{(j,k)}\right). \quad (3)$$

The gate input concatenates current and next states from both models with the normalized token position:

$$g_t = [\widetilde{h}_t^{(1)}; \widetilde{h}_{t+1}^{(1)}; \widetilde{h}_t^{(2)}; \widetilde{h}_{t+1}^{(2)}; p_t]. \quad (4)$$

Layer normalization prevents a high-magnitude layer from dominating the uniform mean.

3.4 Sequence-Level FiLM Conditioning

Inspired by Basani and Chen (2026), we compute the aggregated features from each model's token log-likelihood sequences: mean, standard deviation, variance, skewness, kurtosis, first-difference statistics, second-order variance and entropy, and lag autocorrelation, as well as DFT total energy (Luo et al., 2026). We also append cross-model features including covariance, correlation, cosine similarity, and a Binoculars score (Hans et al., 2024). Let the resulting global vector be s. Inspired by FiLM (Perez et al., 2018), PAWN++ computes

$$\hat{s} = \mathrm{LayerNorm}(s), \quad (5)$$

$$[\gamma, \beta] = W_{\text{film}}\hat{s} + b_{\text{film}}, \quad (6)$$

$$F' = (1+\gamma) \odot F + \beta. \quad (7)$$

γ and β are initialized with zeros, so the initial model satisfies $F' = F$ and begins as the unconditioned detector. The resulting affine conditioning refines the complete representation without replacing its token pathway.

Figure 2 shows the high-level schema of the complete PAWN++ model. Both LLMs remain frozen; only the lightweight networks are optimized.

4 Experimental Setup

4.1 Data

We use the MAGE cross-domain, cross-model split (Li et al., 2024), reversing the original labels to match PAWN: 0 denotes human and 1 denotes machine. The full training, validation, and test sets contain 319,071, 56,792, and 60,743 examples, respectively. The test set is nearly balanced (30,478 human and 30,265 machine). Development ablations use balanced subsets of 5,000 training and 2,000 validation examples; every reported sampled-regime score is still measured on the full 60,743-example test set.

OOD evaluation uses MAGE's unseen-domain-and-model GPT split and its paraphrasing-attack

	RAID	HC3		HC3-Chinese		Beemo	
Method	R@5%FPR	AUROC	Avg. R.	AUROC	Avg. R.	H vs. M	H vs. G
RADAR	70.89	54.59	64.38	35.44	47.93	51.41	60.16
Binoculars	80.61	99.30	91.01	97.35	**93.29**	**83.90**	**78.15**
PAWN++ (ours)	**91.83**	**99.56**	**93.91**	**97.98**	91.96	81.85	77.78

Table 3: One shot transfer of MAGE trained checkpoints. RAID reports average recall at a five percent false positive budget over all generator and decoding groups, HC3 reports AUROC and average recall, and Beemo reports AUROC on human versus machine and human versus GPT-4o-edited text. Full per setting tables are in Appendix B.

split. These testbeds isolate transfer to a new generator and robustness to post-editing.

4.2 Models and Training

The primary and secondary frozen probes are Llama-3.2-1B-Instruct and Llama-3.2-1B (Grattafiori et al., 2024). Architecture and optimization hyperparameters are listed in Appendix A. The best checkpoint is selected by validation AUROC.

The project includes single-model PAWN and fine-tuned RoBERTa-base baselines. For the final comparison tables, we instead use the original PAWN article's published MAGE baselines and report AUROC, class-specific recall, and their arithmetic mean.

5 Results on MAGE

5.1 Ablation Study

Table 1 reports the ablations trained on a balanced subset of 5,000 MAGE training examples and selected using a balanced 2,000-example validation subset. All configurations use the same subsets and are evaluated on the full 60,743-example test set. The original PAWN reaches 0.8753 AUROC with the Instruct probe (Llama-3.2-1B-Instruct) and 0.8953 with the Base probe (Llama-3.2-1B). Adding the second model's token metrics raises AUROC to 0.9126. XPPL and standard hidden-state fusion reduce this score to 0.9096 and 0.9045, respectively, whereas uniform layer fusion raises it to 0.9288. The complete PAWN++ reaches the highest AUROC, 0.9309, after adding sequence-level conditioning. Thus, second-model token metrics provide the initial cross-model gain, uniform fusion is the strongest architectural recovery, and global features add a smaller final improvement.

5.2 In-Distribution MAGE Test

Table 2 compares PAWN++ with the MAGE Testbed 4 baselines reported in the original PAWN article (Miralles-González et al., 2026). PAWN++ reaches 98.36% AUROC and the highest average recall in this comparison (95.16%). Its AUROC is competitive with the strongest reported PAWN and Longformer results, while its class recalls are more balanced than the Longformer baseline.

We additionally evaluate PAWN++ on the two in-distribution settings reported in Basani and Chen (2026): Testbed 2 uses arbitrary domains with the fixed generator GPT-J, while Testbed 3 uses the fixed WritingPrompts (WP) domain with arbitrary generators. PAWN++ reaches 99.94% AUROC and 98.89% average accuracy on Testbed 2, and 99.81% AUROC and 98.72% average accuracy on Testbed 3. In both settings, it exceeds the reported DivEye AUROC and average accuracy.

5.3 Out-of-Distribution Evaluation

Table 8 places the PAWN++ OOD results beside the original PAWN article's MAGE Testbeds 7 and 8 baselines (Miralles-González et al., 2026). On the unseen-domain-and-model testbed, PAWN++ has the highest AUROC and average recall.

Paraphrasing produces a qualitatively different outcome. PAWN++ has 79.13% AUROC, which is below RADAR-FT's 84.00%, confirming that paraphrase robustness remains unresolved. Looking more closely PAWN++ retains high human recall (0.9304) but AI recall falls to 0.4625, indicating that many paraphrased generations cross the fixed decision boundary. Its 79.13% AUROC trails RADAR-FT (84.00%), although it exceeds both original PAWN variants on this metric. The attack therefore removes or weakens the likelihood regularities on which the richer model relies; architecture alone does not substitute for paraphrase-aware training and threshold calibration.

6 One Shot Transfer

A detector that needs retraining for every target distribution cannot serve as shared infrastructure.

We therefore apply the MAGE trained PAWN++ checkpoint to three external benchmarks exactly as it is, with no weight updates, and decision thresholds follow each benchmark's published evaluation protocol. Table 3 summarizes the transfer, and Appendix B carries the complete comparisons.

6.1 RAID

RAID evaluates detectors by recall at a fixed five percent false positive rate across eleven generators, four decoding configurations, and adversarial attacks (Dugan et al., 2024). Table 11 places PAWN++ in the published comparison. With per domain thresholds held at the target budget, PAWN++ reaches 91.83 percent average recall across all generator and decoding groups, ahead of every detector in the comparison that was not trained on RAID, including Binoculars at 80.61 and the commercial systems. The margin is widest exactly where zero shot methods break, on sampling with repetition penalty, where PAWN++ keeps recall above 92 percent while Binoculars falls to 0.6 percent on non chat models.

6.2 Cross-Dataset Evaluation on HC3

HC3 is a bilingual question-answering benchmark that pairs human and ChatGPT answers to the same questions across multiple domains. Its English and Chinese releases provide distinct cross-lingual evaluation settings for machine-generated-text detection (Guo et al., 2023).

We evaluate PAWN++ trained on the full MAGE training set directly on both HC3 versions, without adaptation. On HC3 English, PAWN++ achieves 99.56 AUROC and 93.91 average recall, the strongest results in Table 10. On HC3-Chinese, it reaches 97.98 AUROC and 91.96 average recall (Table 10), the highest AUROC in the comparison, while Binoculars keeps the best average recall at 93.29.

6.3 Cross-Dataset Evaluation on Beemo

Beemo is a multi-author machine-generated-text benchmark containing human-written responses, outputs from ten instruction-tuned LLMs, expert-edited outputs, and outputs edited by Llama3.1-70B-Instruct or GPT-4o. It covers open-ended generation, rewriting, summarization, and open and closed question answering, making it suitable for evaluating both cross-dataset transfer and robustness to post-generation editing (Artemova et al., 2025).

We evaluate the PAWN++ checkpoint trained on the full MAGE training set directly on Beemo, without fine-tuning or threshold calibration on Beemo. Table 9 adds PAWN++ to the seven binary AUROC comparisons reported by Artemova et al. (2025). PAWN++ reaches 81.85 AUROC on human-written versus unedited machine-generated text and remains strong when Llama- or GPT-edited text is contrasted with human writing (81.82 and 77.78, respectively). Expert editing remains challenging: the human-versus-expert-edited AUROC is 63.95. The near-random Llama-edited-versus-machine score (51.87) further indicates that PAWN++ tends to retain a machine-generated signal after Llama editing, but does not reliably distinguish the edited output from its unedited machine-generated counterpart.

7 WATTLE: A Web Benchmark

Existing benchmarks draw human text from curated sources, while web pages carry boilerplate and noisy register, arrive in many languages, and machine text on the web is produced by the strongest available generators. We build **WATTLE** (Web Authorship Testbed for Transformations, Languages, and Eras), a benchmark of 41,300 objects in which every machine text is anchored to a real web page and every comparison is source matched. Figure 3 maps the composition, and Appendix C details every cell (Table 6).

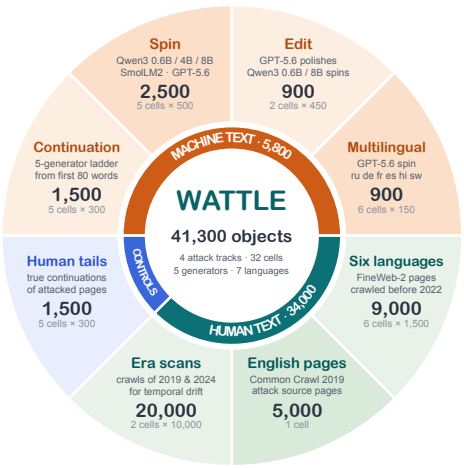


Figure 3: Composition of WATTLE. Four machine tracks anchored to source pages, three human blocks predating the modern generation era, and matched tail controls.

Human gold. All human text predates the post ChatGPT proliferation of machine content. English pages come from a 2019 Common Crawl snapshot (Common Crawl Foundation, 2019), and six further languages (ru, de, fr, es, hi, sw) come from FineWeb-2 shards crawled before 2022 (Penedo et al., 2025). Pages are deduplicated globally, capped per host, and assigned to calibration and test partitions by a salted hash of the URL. Two era scans of ten thousand pages each, drawn from 2019 and 2024 crawls, support a temporal analysis.

Tracks. Machine text is generated from the same source pages along four tracks. *Spin* rewrites the page in the generator's own words across a capability ladder of Qwen3 0.6B to 8B (Qwen Team, 2025), SmolLM2 1.7B (Ben Allal et al., 2025), and the frontier model GPT-5.6. *Continuation* extends the page from its first eighty words, with matched human tail controls that pit mid document machine text against the true continuation of the same page. *Edit* has the frontier model lightly polish machine rewrites from a weak and a strong generator (Artemova et al., 2025). *Multilingual* applies frontier spin in all six languages against same language human pages.

Protocol. Generation is one shot, with one fixed prompt per track, fixed sampling, one attempt per seed, and no regeneration of rejected outputs. Every generation passes a single acceptance predicate covering length, refusal filters, language identity, and bounded source overlap. Released cells are subsampled to round sizes by ascending hash of the URL, a rule fixed before any detector was run, so membership cannot leak detector information. The release ships a manifest with per file checksums, full provenance, and complete acceptance statistics (Appendix C).

8 Can Detectors Filter the Web?

8.1 Methods and Protocol

We compare seven frozen detector families: PAWN++, canonical Binoculars with the Falcon-7B pair (Hans et al., 2024; Almazrouei et al., 2023), a probe matched perplexity ratio sharing the Llama probes of PAWN++, Fast-DetectGPT (Bao et al., 2024), a plain log likelihood score, RADAR (Hu et al., 2023), FastText (Joulin et al., 2017), and RoBERTa fine tuned on MAGE (Liu et al., 2019). Every machine text is paired with the human score of its own source page, operating points use each method's one percent threshold from a held out calibration split, uncertainty comes from bootstrap over URL clusters, and method deltas are tested with McNemar on shared pages. A host disjoint recomputation of the English tracks moves AUROC by at most 0.30 points for every method (Appendix C).

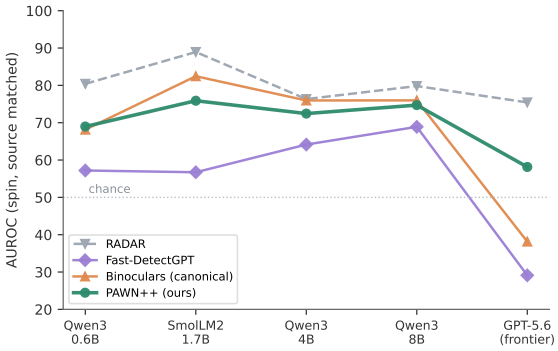


Figure 4: Spin detectability across the generator ladder. Every frozen likelihood detector weakens toward the frontier generator, ratio and curvature scores fall below chance, and PAWN++ stays above chance while degrading gracefully.

8.2 Results

Continuation from open generators is detectable. PAWN++ reaches 97.9 to 100.0 AUROC per open generator cell with 81 to 99 percent recall at the one percent budget, and matched tail controls confirm the signal survives a mid discourse comparison. Supervised baselines rank the classes well and collapse at the operating point.

The frontier regime breaks frozen detectors. Rewrites by GPT-5.6 invert likelihood scores. Canonical Binoculars falls to 38.2 AUROC, the probe matched ratio to 45.6, and Fast-DetectGPT to 29.1, all below chance, while frozen PAWN++ stays above chance at 58.1 (Figure 4). The same shift appears in continuation, where every frozen method drops from near perfect scores to between 45 and 70 AUROC on the frontier cell. Detectability is flat across open generator sizes from 0.6B to 14B, and the frontier generator forms a distinct low detectability outlier. No frozen method recalls more than 12 percent of frontier rewrites at the one percent budget.

Calibrated budgets fail to transfer across languages. Applying the English threshold to six languages breaks every frozen likelihood method somewhere. The probe matched ratio reaches 9.5

	Continuation		Spin		Frontier spin
Method	AUROC	TPR@1%	AUROC	TPR@1%	AUROC
Log likelihood	64.4	6.5%	55.4	0.5%	44.7
Fast-DetectGPT	67.3	26.7%	53.3	2.8%	29.1
FastText	73.5	14.6%	65.8	4.3%	60.9
RoBERTa (MAGE)	87.5	27.5%	73.5	6.9%	67.3
RADAR	94.8	20.3%	80.3	1.5%	75.4
probe matched ratio	94.7	77.9%	77.7	11.8%	45.6
Binoculars	93.0	75.2%	67.9	11.7%	38.2
PAWN++ (ours)	92.0	**74.8%**	70.1	8.6%	**58.1**

Table 4: Detection on WATTLE by regime. Source matched paired AUROC and recall at each method's own calibrated one percent false positive threshold. Continuation and spin pool five generator cells each, and frontier spin is the GPT-5.6 rewrite cell. Multilingual results with backbone coverage conditioning are in Appendix E.

percent false positives on Hindi, a language its backbone officially supports, and 65 percent on Swahili outside its coverage. RADAR and FastText keep their budgets with zero recall. Frozen PAWN++ degrades least, with a worst case of 1.9 percent on supported languages (Appendix E).

Temporal drift. The flagged fraction under the frozen threshold rises from 1.49 percent on the 2019 crawl to 13.44 percent on 2024 (Appendix F), observational evidence of a changing web.

8.3 Training on the Benchmark Inflates Scores and Breaks Transfer

PAWN++ trains a 2.7M parameter head over frozen probes, so supervising it for a new target costs one epoch on a few thousand texts. We train a fresh head on five thousand balanced texts from 2024 crawl pages whose URLs never appear in the benchmark, with the machine side produced by the open generator wave of the same pipeline, and evaluate on the frozen release, which is disjoint from training in URLs and time.

The in protocol numbers look spectacular. Pooled spin rises from 70.1 to 96.4 AUROC, continuation reaches 99.4, the held out frontier cell rises from 58.1 to 96.8, and the six held out languages score 90.5 to 96.1 with false positive rates at or below 2.7 percent. A FastText classifier trained on the same texts reaches only 56.9 AUROC on the frontier cell, so these gains are not a surface artifact that a lexical model could copy.

External evaluation reverses the verdict. On Beemo, machine text produced by a different pipeline with different prompts, the benchmark trained head scores 10 to 26 AUROC points below the frozen head on every comparison, falling from 84.5 to 58.5 on unedited machine text. Against matched human tails the benchmark trained continuation numbers drop from 99.4 to 80.0. The head therefore learned the generation protocol of the benchmark rather than machine text in general, and its benchmark scores overstate real capability by a wide margin. We report the frozen configuration as the honest measure of every detector throughout this paper, and we read this experiment as a warning for the field: a detector trained on a benchmark cannot be certified by that benchmark, and deployment grade evaluation needs generation pipelines that the detector never saw.

9 Conclusion

This work carried a single detector from a curated benchmark into deployment conditions. On MAGE, PAWN++ delivers the most balanced in distribution results in its comparison and the best transfer under domain and generator shift, and the untouched checkpoint holds up across RAID, HC3, and Beemo. The picture changes on WATTLE: statistics that looked reliable for years stop separating classes once the generator is a frontier model, and thresholds calibrated on English quietly lose their meaning in other languages. A closing experiment trains the small head on five thousand texts from the benchmark pipeline and finds that the resulting scores are an illusion: in protocol AUROC jumps to 96.4 while the frozen checkpoint stays ahead by up to 26 points on external machine text. Detection on the web remains open, and progress claims should rest on frozen evaluation against generation pipelines the detector never saw.

Limitations

Our experiments use compact 1B-parameter probe models, so the reported behavior is specific to this compute and model regime. WATTLE covers six non-English languages and uses one fixed prompt

per transformation track and a single frontier-model provider. Its results therefore do not establish generalization to other languages, prompt families, or providers. Generator behavior changes rapidly, so the findings characterize only the documented model versions and evaluation period. The temporal crawl is unlabeled, and changes in the fraction of pages flagged by a detector cannot be interpreted as estimates of AI-generated content prevalence. Multilingual operating points are calibrated on English and depend on the language coverage of the frozen probes. Finally, detector scores provide statistical evidence about text provenance and should not be treated as standalone proof of authorship for an individual document.

Ethics Statement

All human text in WATTLE originates from publicly available web crawls released for research use, Common Crawl (Common Crawl Foundation, 2019) and FineWeb-2 (Penedo et al., 2025). We do not intentionally collect private data or target individual authors, and public pages may still contain personal information, which the release mitigates by storing only what reproducibility requires. Pages serve as corpus documents, the release stores only the URLs and text needed for reproducibility, and no attempt is made to identify or profile the people behind them. Machine text is generated by publicly documented models under their respective licenses. Detection scores are intended to support moderation and corpus curation with human oversight, and this work does not develop or release techniques for evading detection. All experiments used shared academic infrastructure with compact 1B parameter probes, keeping the compute footprint modest.

Acknowledgments

This work is an output of a research project (HSE-BR-2025-025) implemented as part of the Basic Research Program at HSE University.

References

P. Akhilesh, Anjali Chennupati, Bhamidipati Prahas, Bommisetty Durga Jasvitha, and Manju Venugopalan. 2025. AI generated text detection using LORA based fine-tuning of LLM models. In *2025 3rd World Conference on Communication & Computing (WCONF)*, pages 1–7. IEEE.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Ekaterina Artemova, Jason S. Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. 2025. Beemo: Benchmark of expert-edited machine-generated outputs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6992–7018, Albuquerque, New Mexico. Association for Computational Linguistics.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Advik Raj Basani and Pin-Yu Chen. 2026. Diversity boosts AI-generated text detection. *Transactions on Machine Learning Research*. ArXiv:2509.18880.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. 2025. SmolLM2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Common Crawl Foundation. 2019. The Common Crawl corpus. `https://commoncrawl.org`.

Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. The impact of AI-generated text on the internet. *Preprint*, arXiv:2604.26965.

George Drayson, Emine Yilmaz, and Vasileios Lampos. 2025. Machine-generated text detection prevents language model collapse. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29657–29673, Suzhou, China. Association for Computational Linguistics.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Arka Dutta, Utkarshani Jaimini, Utkarsh Bhatt, Sara Shree Muthuselvam, Amitava Das, and Ashiqur Rahman KhudaBukhsh. 2026. A large scale

social web audit of AI generated text detection systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 677–690.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17519–17537. PMLR.

Sichang Steven He, Calvin Ardi, Ramesh Govindan, and Harsha V. Madhyastha. 2026. DeGenTWeb: A first look at LLM-dominant websites. *Preprint*, arXiv:2605.00087.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-text detection via adversarial learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 15077–15095.

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, Valencia, Spain. Association for Computational Linguistics.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Haitong Luo, Weiyao Zhang, Suhang Wang, Wenji Zou, Chungang Lin, Xuying Meng, and Yujun Zhang. 2026. SpecDetect: Simple, fast, and training-free detection of LLM-generated text via spectral analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32356–32364.

Pablo Miralles-González, Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2026. Not all tokens are created equal: Perplexity attention weighted networks for AI-generated text detection. *Information Fusion*, 125:103465.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Marco Murgia, Diego Reforgiato Recupero, and Georgios Spathoulas. 2026. A comprehensive evaluation of fine-tuned LLMs as AI-generated text detectors under adversarial and multi-domain conditions. *IEEE Access*, 14:66819–66837.

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 pretraining data processing to every language. *Preprint*, arXiv:2506.20920.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3942–3951.

Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, and Tianxing He. 2023. On the zero-shot generalization of machine-generated text detectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4799–4808, Singapore. Association for Computational Linguistics.

Gerrit Quaremba, Elizabeth Black, Denny Vrandečić, and Elena Simperl. 2025. WETBench: A benchmark for detecting task-specific machine-generated text on Wikipedia. *Preprint*, arXiv:2507.03373.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Brian Tufts, Xuandong Zhao, and Lei Li. 2025. A practical examination of AI-generated text detectors for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4839–4856, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaowei Zhu, Yubing Ren, Fang Fang, Shi Wang, Yanan Cao, and Li Guo. 2026. Exons-Detect: Identifying and amplifying exonic tokens via hidden-state discrepancy for robust AI-generated text detection. *Preprint*, arXiv:2603.24981.

A Training Hyperparameters

Table 5 lists the architecture and optimization settings shared by every PAWN++ training run in this paper.

Hyperparameter	Value
Maximum sequence length	512
MLP width	256
MLP hidden layers	3
Optimizer	AdamW
Maximum epochs	5
Learning rate	10^{-3}
Weight decay	10^{-2}
Token dropout	0.15
Label smoothing	0.2
Batch size	32
Random seed	42
Checkpoint criterion	Validation AUROC

Table 5: PAWN++ architecture and training hyperparameters.

B Complete Benchmark Results

This appendix reports every published comparison in full for the benchmarks of Sections 5 and 6.

Table 7 reports the two additional in distribution MAGE settings, a fixed generator with arbitrary domains and a fixed domain with arbitrary generators, where PAWN++ exceeds every published baseline including DivEye.

Table 8 places PAWN++ beside the published out of distribution baselines on MAGE testbeds 7 and 8. The unseen domain and model testbed shows the strongest transfer in the comparison, and the paraphrase testbed shows the operating point cost that motivates the benchmark training study of Section 8.3.

Table 9 gives all seven binary Beemo comparisons against the published baselines, covering unedited machine text and machine text refined by experts or stronger models.

Table 10 reports the complete metric sets on HC3 English and Chinese side by side for checkpoints applied without adaptation.

Table 11 expands the RAID comparison into the full recall grid over generator groups, decoding strategies, and repetition penalty, with the published values reproduced from prior work.

C The WATTLE Benchmark

Table 6 breaks the release into its 32 cells. The machine side holds 5,800 texts across four attack tracks, every one generated from a source page that also appears on the human side. The generator ladder spans open weight models from 0.6B to 8B parameters and one frontier model, so the benchmark measures detectors against both accessible and strongest available generation. The human side holds 34,000 documents, and the five control cells hold the true continuations of the pages attacked in the continuation track.

Sources. English human gold comes from the CC-MAIN-2019-30 snapshot of Common Crawl (Common Crawl Foundation, 2019) with boilerplate removal, global URL deduplication, and per host caps. The six language cells use FineWeb-2 shards crawled before 2022 (Penedo et al., 2025) with 1,500 pages per language. The released human side spans 29,386 unique hosts across its 34,000 pages, so no single site dominates the benchmark. A salted URL hash assigns every document to the calibration or test partition.

Generation prompts. Each track uses one fixed prompt for every generator, with default sampling for the frontier model and temperature 1.0 with top-p 0.95 for the open weight ladder. The spin track uses:

Spin Prompt.

```
Rewrite the following web page in your own words. Keep the same topic, the same language, roughly the same length, and the same level of detail. Output only the rewritten text, with no preamble or commentary.
```

The continuation track receives the first eighty words of the page and a length target derived from the original:

Continuation Prompt.

```
Continue the following web page. Keep the same topic, language, style and register. Write about {target words} more words. Output only the continuation, with no preamble or commentary.
```

The edit track passes a machine rewrite back to the frontier model:

Edit Prompt.

```
Lightly edit the following text to improve its flow and wording. Fix awkward phrasing where needed, but keep the meaning, structure, language, and approximately the same length. Output only the edited text, with no preamble or commentary.
```

The multilingual track reuses the spin prompt, and the phrase about keeping the same language constrains the output to the language of the source page.

Acceptance and release. The acceptance predicate requires at least 400 characters, passes refusal and preamble filters, verifies language identity, and bounds source overlap at 0.5 for rewrites within a length ratio of 0.6 to 1.6, while edits must keep

Track	Cells	Sources and generators	Per cell	Total
Machine text, anchored to source pages				
Spin	5	Qwen3 0.6B/4B/8B, SmolLM2 1.7B, GPT-5.6	500	2,500
Continuation	5	the same generator ladder	300	1,500
Edit	2	GPT-5.6 over Qwen3 0.6B and 8B rewrites	450	900
Multilingual	6	GPT-5.6 in ru, de, fr, es, hi, sw	150	900
Human text, pre LLM				
English pages	1	Common Crawl 2019	5,000	5,000
Per language	6	FineWeb-2 crawled before 2022	1,500	9,000
Era scans	2	crawls of 2019 and 2024	10,000	20,000
Controls				
Human tails	5	true continuations of source pages	300	1,500
WATTLE total	**32**			**41,300**

Table 6: Composition of WATTLE. Every machine text is anchored to a real web page, released cells are selected from accepted generation pools by a score blind hash rule, and human text predates the post ChatGPT era.

Testbed 2: Arbitrary domains, GPT-J

Method	AUROC	Human R.	AI R.	Avg. R.
LogRank	67.00	58.81	63.94	61.38
Entropy	83.00	76.43	76.84	76.64
DetectLLM	72.00	66.36	62.07	64.21
FastDetectGPT	59.00	62.31	50.49	56.40
Binoculars	69.00	60.11	65.22	62.67
BiScope	93.00	89.62	84.86	87.24
DivEye	97.00	90.63	88.56	89.60
PAWN++ (ours)	**99.94**	**99.24**	**98.54**	**98.89**

Testbed 3: WritingPrompts, arbitrary models

Method	AUROC	Human R.	AI R.	Avg. R.
LogRank	76.00	89.61	56.15	72.88
Entropy	78.00	85.96	60.40	73.18
DetectLLM	91.00	88.54	80.77	84.66
FastDetectGPT	76.00	87.25	54.08	70.67
Binoculars	77.00	80.80	62.07	71.44
BiScope	94.00	91.78	95.27	93.53
DivEye	99.00	92.22	96.88	94.55
PAWN++ (ours)	**99.81**	**98.97**	**98.47**	**98.72**

Table 7: Additional MAGE in-distribution results. Baselines, testbed definitions, and DivEye values are from Basani and Chen (2026); PAWN++ is our result in the corresponding MAGE settings.

Testbed 7: Unseen domain and model

Method	AUROC	Human R.	AI R.	Avg. R.
Longformer	94.00	52.50	**99.14**	75.82
Longformer†	94.00	88.78	84.12	86.54
RADAR*‡	87.00	82.41	73.62	78.02
RADAR-FT	94.00	61.42	98.12	79.77
Binoculars*‡	88.00	5.51	7.12	17.20
LogRank	68.00	83.87	43.95	63.91
Entropy	66.00	74.93	50.18	62.55
DetectLLM	73.00	63.66	67.40	65.53
FastDetectGPT	70.00	93.38	41.50	67.44
BiScope	92.00	86.00	82.58	84.24
DivEye	87.00	69.75	83.22	76.49
PAWN (GPT2)	94.00	84.12	89.50	86.81
PAWN (Llama3-1B)	97.00	78.87	96.75	87.81
PAWN++ (ours)	**98.42**	**93.04**	95.38	**94.21**

Testbed 8: Paraphrasing attacks

Method	AUROC	Human R.	AI R.	Avg. R.
Longformer	75.00	52.16	81.73	66.94
Longformer†	75.00	88.78	37.05	62.92
RADAR*‡	71.00	67.67	59.87	63.77
RADAR-FT	84.00	53.33	**92.25**	72.79
Binoculars*‡	55.00	6.27	16.00	21.76
BiScope	81.00	48.80	89.79	69.30
DivEye	**87.00**	69.75	83.22	**76.49**
PAWN (GPT2)	75.00	87.00	39.13	63.06
PAWN (Llama3-1B)	74.00	77.53	55.87	66.70
PAWN++ (ours)	79.13	**93.04**	46.25	69.65

Table 8: OOD MAGE results. Original baselines are from Miralles-González et al. (2026); additional methods are from Basani and Chen (2026); PAWN++ is our result. The additional source reports class accuracies (HumanAcc/MachineAcc) and AvgAcc. * unsupervised; ‡ threshold not selected; † threshold selected for Longformer.

overlap between 0.15 and 0.95. Acceptance statistics ship with the release manifest together with per file checksums, prompts, and generator versions. The split is URL level. Re evaluating the English tracks on test pages whose host never occurs in calibration changes AUROC by at most 0.30 points for every method, so host overlap does not affect the results.

D Per Cell Results and Probe Order Control

Table 12 breaks the pooled web results of Table 4 into representative cells and adds the reversed probe order of Binoculars as a control.

Matched human tail controls give 95.4 to 98.9 AUROC for PAWN++ on open generator continuations, so continuation detection is a signal about the text rather than its position in the document. On

Detector	H vs. M	E vs. M	H vs. E	L vs. M	H vs. L	G vs. M	H vs. G
Training-free detectors							
Binoculars	83.90	**76.79**	61.24	59.90	79.90	57.93	78.15
Log Probability	**87.77**	68.68	**78.02**	56.23	**84.52**	62.45	**81.27**
Rank	74.22	57.65	69.11	52.55	72.52	55.10	70.81
Log-Rank	86.91	68.13	77.04	56.04	83.62	62.09	80.33
Entropy	10.36	32.02	17.74	44.01	12.82	38.01	15.28
DetectLLM: Likelihood Log-Rank Ratio	79.68	64.12	69.40	54.71	76.25	59.41	72.83
DetectLLM: Normalized Perturbed Log-Rank	74.12	68.30	68.11	54.52	72.49	50.68	73.84
DetectGPT	72.76	69.37	68.21	61.36	70.14	63.05	69.90
Training-based detectors							
FastText	60.18	56.09	54.24	57.34	53.03	54.47	55.39
RADAR	51.41	50.93	50.80	34.51	62.65	37.96	60.16
MAGE	73.72	60.88	64.09	57.95	67.79	60.59	65.37
AIGC MPU	70.52	70.35	50.66	**65.71**	57.17	59.66	62.69
OpenAI RoBERTa-base	66.95	66.83	49.13	66.95	50.00	**66.95**	50.00
OpenAI RoBERTa-large	62.93	63.96	48.76	64.71	48.57	59.67	53.74
PAWN++ (ours)	81.85	70.14	63.95	51.87	81.82	57.13	77.78

Table 9: AUROC scores (%) on Beemo. Bold denotes the best score in each column; underlining denotes the best training-based score. Published baselines and task definitions are from Artemova et al. (2025). PAWN++ and the 50-epoch FastText model are evaluated using checkpoints trained on full MAGE, without adaptation to Beemo. For all single-model training-free detectors, the Qwen2-7B configuration is reported. H=human-written; M=machine-generated; E=expert-edited; L=Llama3.1-70B-Instruct-edited; G=GPT-4o-edited. L and G aggregate three editing prompts.

Method	AUROC	Human R.	AI R.	Avg. R.
RoBERTa-base	98.04	60.87	99.14	80.00
Binoculars	99.30	82.68	99.33	91.01
RADAR	54.59	34.11	94.66	64.38
Fast-DetectGPT	99.41	83.71	99.15	91.43
FastText	68.65	55.69	74.60	65.14
PAWN++ (ours)	**99.56**	**88.29**	**99.54**	**93.91**

Method	AUROC	Human R.	AI R.	Avg. R.
RoBERTa-base	53.55	30.36	71.79	51.08
Binoculars	97.35	92.37	94.21	**93.29**
RADAR	35.44	92.33	3.53	47.93
Fast-DetectGPT	97.55	89.34	93.83	91.58
FastText	53.79	**98.90**	5.02	51.96
PAWN++ (ours)	**97.98**	87.89	**96.04**	91.96

Table 10: One shot transfer to HC3 English answers (left) and Chinese answers (right) for checkpoints applied without adaptation.

a shared URL core attacked by all five spin generators, PAWN++ moves from 71.9 to 79.3 across the open ladder and dips to 68.9 on the frontier cell, while the probe matched ratio falls from 84.4 to 58.6. An observational extension on a 2024 generation wave adds a Qwen3-14B point and shows flat detectability from 0.6B to 14B. The largest single drop in the comparison belongs to the probe matched ratio, which loses 26 AUROC points between the shared core and the frontier cell.

E Multilingual Results and Backbone Coverage

Interpreting multilingual results requires the language coverage of each backbone. Llama-3.2 officially supports de, fr, es, and hi among our six, Russian is present in its web pretraining without being listed, and Swahili lies outside both. Falcon-7B covers English and French. RADAR, FastText, and RoBERTa are English only, and RoBERTa is not evaluated on non English text by design. Cells outside a backbone's coverage measure pretraining exposure rather than detection and are excluded from headline claims. Figure 5 plots the per language AUROC of every frozen method, and Table 13 adds the operating point detail. Fast-DetectGPT and the probe matched ratio break the transferred budget on Hindi and Swahili, while PAWN++ keeps the realized false positive rate at or below 1.4 percent on every supported language.

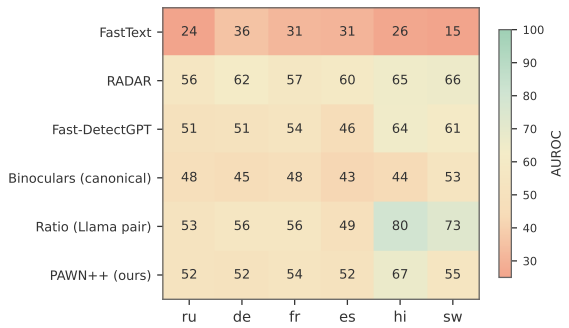


Figure 5: Per language AUROC on frontier multilingual spin. Frontier spin stays close to chance for every frozen method, and cells outside a backbone's pretraining coverage measure exposure rather than detection.

Table 11: Recall at 5% FPR on RAID (Dugan et al., 2024).

	Open Source								Closed Source				Avg.
	Chat (llama-c, mistral-c, mpt-c)				No Chat (mistral, mpt, gpt2)				Chat (c-gpt, gpt4, cohere)		No Chat (cohere, gpt3)		
Decoding strategy	Greedy		Sampling		Greedy		Sampling		Greedy	Sampling	Greedy	Sampling	
Repetition penalty	No	Yes	No	Yes	No	Yes	No	Yes	No	No	No	No	
No training on RAID													
R-B GPT2	84.10%	52.30%	77.90%	26.20%	98.60%	44.10%	60.50%	35.40%	70.90%	41.70%	65.10%	52.50%	59.11%
R-L GPT2	79.70%	41.10%	71.40%	19.50%	98.50%	43.00%	67.20%	53.40%	61.40%	34.70%	61.10%	48.60%	56.63%
R-B CGPT	80.20%	63.30%	75.50%	39.30%	53.30%	26.40%	14.90%	1.70%	59.10%	38.10%	46.50%	39.00%	44.78%
RADAR	88.80%	77.40%	85.60%	66.40%	91.80%	63.80%	48.30%	31.80%	81.60%	75.30%	72.20%	67.70%	70.89%
GLTR	89.80%	67.50%	83.90%	38.30%	99.60%	56.90%	44.50%	0.50%	80.70%	54.30%	75.60%	63.70%	62.94%
F-DetectGPT	98.60%	74.50%	96.20%	40.50%	97.80%	56.10%	79.70%	0.60%	96.00%	74.10%	93.80%	86.30%	74.52%
LLMDet	55.50%	30.20%	47.50%	16.50%	74.80%	27.00%	38.40%	3.70%	35.80%	18.50%	40.00%	32.90%	35.07%
Binoculars	99.90%	86.60%	99.70%	60.60%	99.90%	62.30%	72.40%	0.60%	99.20%	92.10%	99.00%	95.00%	80.61%
GPTZero	98.80%	93.70%	98.40%	82.50%	74.70%	34.60%	9.40%	4.80%	92.30%	88.50%	60.60%	53.40%	65.98%
Originality	98.60%	86.30%	97.70%	72.50%	99.90%	64.10%	89.00%	51.20%	96.80%	89.00%	91.70%	85.40%	85.18%
Winston	97.20%	90.10%	96.60%	78.30%	68.20%	49.00%	29.50%	11.30%	96.10%	93.70%	73.20%	68.10%	70.94%
ZeroGPT*	95.40%	80.70%	90.50%	54.90%	85.10%	57.20%	16.00%	0.30%	92.10%	65.80%	83.40%	72.70%	66.18%
PAWN (GPT2)	96.56%	88.57%	93.70%	79.30%	90.53%	84.03%	57.52%	81.56%	90.16%	72.02%	84.34%	75.37%	82.81%
PAWN (Llama3-1b)	98.33%	92.00%	96.26%	81.61%	98.90%	86.85%	52.07%	82.44%	93.10%	73.24%	83.15%	75.04%	84.42%
Longformer	87.89%	86.60%	86.55%	83.43%	86.15%	83.78%	59.62%	83.63%	81.66%	68.24%	64.13%	54.11%	77.15%
RoBERTa	89.89%	98.10%	91.14%	90.26%	85.32%	95.55%	49.14%	89.24%	89.04%	71.97%	68.53%	66.02%	82.02%
RADAR-PTM	88.14%	87.05%	85.80%	82.54%	94.10%	88.66%	49.15%	92.05%	83.51%	67.96%	66.29%	56.80%	78.50%
PAWN++ (ours)	**99.39%**	**98.28%**	**99.50%**	**92.82%**	**91.90%**	**97.80%**	**69.03%**	**95.33%**	**97.41%**	**84.57%**	**92.23%**	**83.65%**	**91.83%**

Notes. Source values are reproduced from Table 10 of Miralles-González et al. (2026). PAWN++ was evaluated on all 467,985 clean RAID-train rows (Dugan et al., 2024) using a separate threshold for each domain at no more than 5% human FPR. An asterisk marks a source detector that did not achieve the target FPR.

	Spin		Continuation		Edit
Method	Qwen3-8B	GPT-5.6	Qwen3-8B	GPT-5.6	GPT-5.6∘q06
FastText	65.7 / 5%	60.9 / 2%	79.6 / 13%	74.1 / 6%	64.2 / 7%
RoBERTa (MAGE)	82.6 / 13%	67.3 / 3%	94.1 / 34%	88.9 / 9%	78.5 / 7%
RADAR	79.8 / 1%	75.4 / 1%	97.4 / 13%	92.7 / 3%	82.2 / 2%
Fast-DetectGPT	68.9 / 3%	29.1 / 0%	97.7 / 77%	45.5 / 1%	62.5 / 2%
probe matched ratio	87.4 / 9%	45.6 / 0%	100.0 / 99%	73.3 / 8%	85.6 / 8%
Binoculars (Falcon)	76.0 / 9%	38.2 / 0%	100.0 / 100%	70.0 / 11%	72.9 / 4%
PAWN++ (ours)	74.7 / 9%	58.1 / 2%	100.0 / 99%	59.3 / 14%	75.2 / 11%

Table 12: Selected cells, AUROC and recall at the one percent budget, source matched. Reversing the Binoculars probe order collapses spin AUROC to between 34 and 45 on every cell, which confirms the canonical orientation.

Method	ru	de	fr	es	hi	sw
FastText	44.1 / 0.0 / 0°	40.2 / 0.1 / 0°	46.3 / 0.0 / 0°	28.7 / 0.0 / 0°	44.8 / 0.1 / 0°	41.2 / 0.0 / 0°
RADAR	56.1 / 0.1 / 0°	61.7 / 0.0 / 0°	55.2 / 0.1 / 0°	59.3 / 0.1 / 0°	63.7 / 0.1 / 0°	65.8 / 0.0 / 0°
Fast-DetectGPT	50.7 / **6.9** / 5	51.2 / 1.5 / 3	54.2 / 2.1 / 2	45.7 / 2.2 / 0	64.3 / **13.5** / 32	60.8 / **7.7** / 10°
probe matched ratio	54.7 / 0.2 / 2	55.2 / 0.3 / 1	57.0 / 0.6 / 0	48.9 / 0.5 / 1	79.1 / **9.5** / 40	72.8 / **65.0** / 88°
Binoculars	48.2 / 1.5 / 1°	45.1 / 0.1 / 0°	48.0 / 0.5 / 0	42.5 / 0.7 / 0°	43.7 / 0.0 / 0°	53.1 / 0.1 / 0°
PAWN++ (ours)	52.1 / 1.3 / 0	51.9 / 0.2 / 0	53.9 / 1.4 / 5	51.7 / 0.5 / 1	67.1 / 0.2 / 0	54.6 / **6.1** / 5°

Table 13: Multilingual frontier spin per language. Each cell reports AUROC, realized false positive percentage at the English one percent threshold, and recall percentage. The circle marks languages outside the backbone's pretraining coverage. Bold false positive rates mark broken budgets.

F Temporal Analysis

The 2019 scan is flagged at 0.11, 1.49, and 5.81 percent under thresholds calibrated for 0.1, 1, and 5 percent, so the budgets hold on in era human text. The 2024 scan is flagged at 2.64, 13.44, and 18.92 percent. We read the scan as a measure of threshold stability under crawl era shift rather than as an estimate of machine text prevalence, which would require contemporary authorship labels. The widening between the eras appears at every operating point, and the loosest threshold flags nearly one page in five by 2024.

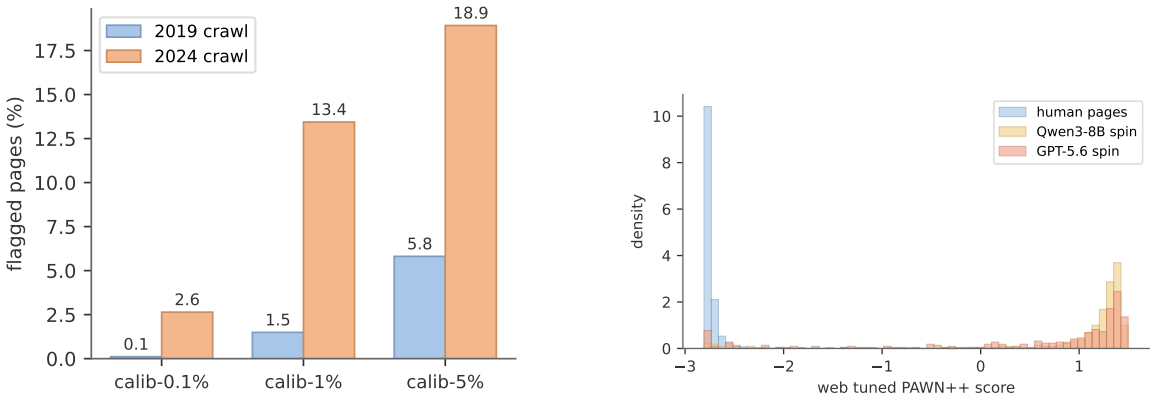


Figure 6: Left: fraction of era scan pages flagged by PAWN++ at three calibrated thresholds, where the 2019 rates match their nominal budgets and the 2024 rates rise far above them. Right: score distributions of the benchmark trained head on human pages, Qwen3-8B spin, and GPT-5.6 spin, whose perfect separation does not survive external machine text (Section 8.3) and is the visual face of benchmark overfitting.

G Selective Prediction

An abstention band built from empirical human score quantiles flags only above the one percent quantile and abstains below the thirty percent quantile, at the cost of abstaining on about thirty percent of human text. For frozen PAWN++ the band lifts continuation recall among decided examples to 100 percent while machine abstention stays at one percent. On spin the same band abstains on about half of the machine texts and decides the rest as human, so abstention rescues continuation and leaves the spin regime open. Audit precision at fifty is measured on a balanced pool and reaches 90 percent on continuation.

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 webpage 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 webpage.

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

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 297–307
October 29, 2026 ©2026 Association for Computational Linguistics

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 Magic-HTML, 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 `5da0972e9b58d0c7891ae75053ced97c268f52e3`. We refer to this revision-specific file as the fixed snapshot. The `webmainbench.jsonl` file has SHA-256 `85765fe798f07c14eb1c92945046eaa56e0da59663f70b9c498647d7dfd78884` 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 HTML $\rightarrow e \rightarrow M_e \rightarrow$ LID and full-page HTML $\rightarrow e \rightarrow F_e \rightarrow$ 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 (Barbaresi, 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

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

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

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

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/n	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]	n
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.level` 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?", *lìjí zhùcè* "Sign up now", *chūcuò le* "Something went wrong", and *qǐng chángshì chóngxīn jiāzài* "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 $G \to M_e$ 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 (OpenDataLab, 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 Pretraining. 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, Jindrich 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.

CEYNEWS: A Sri Lankan Trilingual News Corpus

Ajintha Sivakulasingam[1], Kengatharaiyer Sarveswaran[1], Tharindu Ranasinghe[2]
[1]University of Jaffna, Sri Lanka, [2]Lancaster University, UK
{sajintha,sarves}@univ.jfn.ac.lk, t.ranasinghe@lancaster.ac.uk

Abstract

Language models have achieved state-of-the-art performance across many natural language processing (NLP) tasks, but low-resource languages remain underrepresented due to limited pre-training data and evaluation benchmarks. While recent efforts have introduced corpora, models, and benchmarks for such languages, most focus on a single language in isolation. This is particularly limiting in multilingual countries such as Sri Lanka, where the lack of comparable multilingual resources hinders the development and fair evaluation of models across languages. To address this gap, we introduce CEYNEWS, the largest trilingual news corpus for Sinhala, Tamil, and English. The dataset contains over 1.02 million articles (136 million tokens) collected between 2013 and 2026 from three news sources, with full article text and rich metadata. Using CEYNEWS, we develop three downstream tasks: news source identification, news category classification, and headline generation, and evaluate multilingual models across different learning settings. Our results demonstrate the challenges of applying multilingual models to these tasks, particularly for the headline generation task. CEYNEWS fills a major gap in Sri Lankan and South Asian language resources and is publicly released to support research in multilingual NLP, LLM adaptation, and digital humanities.

1 Introduction

The rapid advancement of language models (LMs) (Devlin et al., 2019; Brown et al., 2020) has transformed natural language processing (NLP), achieving state-of-the-art performance across a wide range of tasks (Ghaseminejad Raeini, 2025). However, the success of LMs is heavily dependent on the availability of large-scale, high-quality pre-training corpora (de Gibert et al., 2024). As a result, their capabilities remain disproportionately concentrated in high-resource languages.

Low-resource languages, in contrast, continue to lag significantly behind (Joshi et al., 2020). This disparity is further amplified by the limited availability of curated benchmarks for low-resource languages, hindering reliable evaluation (Hu et al., 2020; Shardlow et al., 2024a,b). As a result, extending the benefits of language models to low-resource languages remains an open and pressing challenge (Hettiarachchi et al., 2026, 2025).

Addressing this, recent efforts have sought to develop language resources for low-resource languages, including corpora (Hettiarachchi et al., 2024; Charfi et al., 2019), pre-trained models (Dhananjaya et al., 2022; Evdaimon et al., 2024), and benchmarks (Micallef and Borg, 2025; Urbizu et al., 2022). However, most focus on individual languages in isolation, with comparable multilingual resources that enable cross-lingual pre-training and evaluation within a single linguistic or national context remain scarce. This is particularly problematic in multilingual societies, where unequal language representation can lead to weaker models and unfair evaluation across languages (Blasi et al., 2021).

In this work, we address this gap for Sri Lanka, a setting that presents a compelling case for multilingual resource development. In Sri Lanka, Sinhala and Tamil serve as official languages, with English functioning as a link language (Mendis and Rambukwella, 2020). Despite this official multilingualism, large-scale resources spanning all three languages remain scarce (Ranathunga et al., 2025). Sinhala has benefited from recent efforts to develop large-scale corpora (Hettiarachchi et al., 2024), benchmarks (Ranasinghe et al., 2025a,b,c), and language models (Dhananjaya et al., 2022). However, Tamil resources and models are predominantly focused on Indian Tamil (Unjum et al., 2025), resulting in degraded performance on Sri Lankan Tamil (Jeyarajalingam et al., 2026; Mahaganapathy et al., 2026; Thillainathan et al.,

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 308–319
October 29, 2026 ©2026 Association for Computational Linguistics

2026). Although the two varieties are mutually intelligible, they exhibit systematic differences at the phonological, lexical, morphological, and syntactic levels, reflecting their distinct historical development and patterns of language (Unjum et al., 2025; Krishnamurti, 2003; Schiffman, 1999; Kailainathan, 1999).

These challenges highlight a broader limitation in Sri Lankan NLP, as Sinhala primarily suffers from limited data availability, while Sri Lankan Tamil remains underrepresented in existing multilingual resources. Moreover, most datasets, benchmarks, and models have been developed separately for individual languages, resulting in fragmented resources that inadequately support multilingual training and evaluation. This is particularly problematic in Sri Lanka, where public services, education, media, and everyday communication often involve Sinhala, Tamil, and English, with frequent code-switching that monolingual models fail to capture (Ranathunga et al., 2025). Therefore, there is a pressing need for cross-lingual corpora and unified benchmarks that represent all three languages within a shared national context.

In this paper, we introduce CEYNEWS, a unified trilingual news corpus for Sri Lanka, designed to support multilingual and cross-lingual NLP research under realistic national settings. It comprises over 1.02 million articles from three major Sri Lankan outlets-Ada Derana, Hiru, and ITN-spanning 2013-2026. Unlike Sinhala-only corpora such as NSINA (Hettiarachchi et al., 2024) or general-purpose multilingual document collections (Senaratna, 2026), CEYNEWS is purpose-built for benchmarked multilingual and cross-lingual NLP evaluation, providing standardised trilingual splits across Sinhala, Tamil, and English within a single framework. Our **main contributions** are as follows:

(i) We introduce CEYNEWS, the largest trilingual, processable, news corpus for Sri Lanka to date, comprising over 1.02 million articles and 136 million tokens.[1]

(ii) We construct three CEYNEWS-based benchmarks: (i) news source identification, (ii) news category classification, and (iii) headline generation, with standardised train/validation/test splits covering all three languages.

(iii) We evaluate state-of-the-art multilingual models, analysing in-language performance and cross-lingual transfer among the three languages, and provide insights into multilingual generalisation in a low-resource setting.

2 CEYNEWS Corpus

We first describe the corpus collection methodology, followed by a statistical analysis of CEYNEWS.

2.1 Data Collection

We selected three major Sri Lankan news outlets: Ada Derana, Hiru News, and ITN News, based on their coverage of both Sinhala and Tamil. Ada Derana and Hiru News also publish English-language news, enabling the construction of a trilingual corpus, while ITN News was used only for Sinhala and Tamil as no corresponding English-language news section was available.

For each available article, we collected the source, publication timestamp, headline, article content, URL, and news category when the category was available. We developed a separate crawling strategy for each outlet because their structures and content delivery mechanisms differed.

2.1.1 Ada Derana

Ada Derana articles were collected using a sequential, identifier-based crawling strategy. The scraper systematically generated article URLs by incrementing this identifier and retrieved each available article page. Direct HTTP requests to the website encountered a JavaScript-based browser validation mechanism. Therefore, Selenium WebDriver with headless Chrome was initially used to load a seed page and establish a valid browser session. The resulting session cookies were transferred to a persistent `requests.Session`, which was subsequently used to retrieve the public article pages more efficiently. When the session expired, or the browser validation page reappeared, the session–initialisation procedure was repeated. The timestamp, headline, and article content were extracted from each successfully retrieved page using BeautifulSoup. News categories were not collected because this information was not available on the individual article pages. Failed requests were retried using incremental backoff, and a checkpoint containing the most recently pro-

[1]Our dataset is available at https://zenodo.org/records/20579021.

cessed identifier allowed interrupted crawling runs to resume without restarting the collection.

2.1.2 Hiru News

Two crawling strategies were used for Hiru News because its multilingual content was exposed through different interfaces. Sinhala and English articles were retrieved through paginated API endpoints that returned batches of articles in JSON format. The endpoints were queried using the page number and news category. From each response, we extracted the timestamp, headline, article content, URL, and category. Tamil articles were collected from individual article pages containing sequential numeric identifiers in their URLs. The scraper incremented these identifiers, retrieved each available page through HTTP requests, and extracted the timestamp, headline, article content, and category.

2.1.3 ITN News

ITN News articles were collected using a date-based archive traversal strategy. For every date within the predefined collection period, the scraper generated the corresponding daily archive URL and traversed all available pagination pages. Article links were first extracted from these index listings, after which each individual page was retrieved. The timestamp, headline, article content, category, and URL were extracted from each entry and stored in JSON format.

2.2 Data Preprocessing

After data collection, the extracted articles underwent a multi-stage preprocessing pipeline. Unicode NFC normalisation was applied to the headlines and article content, and selected Unicode artefacts introduced during web extraction were removed. Manually defined rule-based cleaning procedures were applied to remove WordPress shortcodes, embedded-media markup, residual HTML elements, URLs, duplicated timestamp headers, separator lines, source attributions, reporter bylines, and website-navigation remnants. Quotation marks and dashes were standardised, and redundant whitespace was collapsed.[2]

Records with missing or empty headlines or article content after preprocessing were excluded. Exact duplicates were identified using the normalized headline–content pair, and only one instance of each duplicate was retained. All available metadata fields were preserved.

2.3 CeyNews Corpus Statistics

After preprocessing, the resulting CeyNews corpus contains approximately 1.026 million articles and 136 million words. We report corpus size in words to provide a model-independent measure that does not depend on the vocabulary or tokenization scheme of a particular language model. For the subsequent token-length analysis, we used the pretrained XLM-RoBERTa tokenizer as a common multilingual subword-tokenization framework (Conneau et al., 2020) because CeyNews is intended to support multilingual NLP and language-modeling applications.

To enhance the historical coverage of CeyNews and avoid redundant data collection, we also provide mappings to relevant portions of the existing NSina corpus (Hettiarachchi et al., 2024). Specifically, Ada Derana Sinhala is mapped to NSina articles published before 2017, while ITN Sinhala is mapped to articles published before 2023. Therefore, the Ada Derana and ITN Sinhala article counts reported in Table 1 include these incorporated NSina articles. The table presents the overall distribution of articles across languages and sources, with Ada Derana and Hiru providing substantial coverage across all three languages.

To provide a more detailed characterization of the linguistic structure of the corpus, we analysed a random sample of 100,000 articles from each language. This analysis was performed on article content only, excluding headlines. Sentence segmentation and word-level tokenization were performed using Stanza[3] for English and Tamil and Sinling[4] for Sinhala. In this analysis, a token refers to a word-level token, and a unique token type refers to a distinct surface word form rather than a lemma. Table 2 reports the resulting sentence, token, vocabulary, and length statistics. The sampled English articles contain 18.64 million word-level tokens across 825,785 sentences, while the corresponding Tamil and Sinhala samples contain 10.09 million and 12.07 million tokens, respectively. English articles contain an average of 186.36 tokens and 8.26 sentences per article. The corresponding averages for Tamil are 100.93 tokens and 7.15

[2]The crawling and preprocessing code is publicly available at https://github.com/LTG-UoJ/CeyNews.

[3]https://stanfordnlp.github.io/stanza/

[4]https://github.com/ysenarath/sinling

Source	Tamil				Sinhala				English			
	From	To	Articles	Tokens	From	To	Articles	Tokens	From	To	Articles	Tokens
Ada Derana	2013	2025	158,289	21,343,134	2008	2025	193,190	26,752,342	2013	2025	78,745	18,086,084
Hiru	2017	2026	154,013	13,174,057	2013	2026	272,275	32,155,375	2013	2026	150,693	25,803,113
ITN	2018	2026	48,671	4,174,477	2022	2026	60,145	8,137,253				
Total			**360,973**	**38,691,668**			**525,610**	**67,044,970**			**229,438**	**43,889,197**

Table 1: Dataset statistics across news sources and languages. The Ada Derana Sinhala counts include NSINA articles published before 2017, while the ITN Sinhala counts include NSINA articles published before 2023 (Hettiarachchi et al., 2024).

sentences per article, while Sinhala articles contain 120.70 tokens and 5.96 sentences per article. At the sentence level, English has an average of 22.57 tokens per sentence, compared with 14.11 for Tamil and 20.26 for Sinhala. The median sentence lengths are 22 tokens for English, 13 for Tamil, and 18 for Sinhala. The sampled Tamil articles contain 514,069 unique surface token types, whereas the English sample contains 214,032 and the Sinhala sample contains 218,946.

For further analysis, we examined XLM-R subword-token length distributions and year-wise article coverage across Sinhala, Tamil, and English. The article distributions were based on data collected by our crawlers between 2013 and 2025. Sinhala articles from Ada Derana before 2017 and ITN before 2023 are excluded from the year-wise analysis, as these articles were obtained from the NSINA corpus. Figures 1 and 2 present the subword-token-length distributions of the headlines and article content, respectively. As shown, most headlines contain 12–20 tokens, whereas most articles contain approximately 75–225 tokens. Sinhala has the largest annual article volume, reaching around 48,000 articles in 2024, followed by Tamil, which peaks at approximately 42,000 articles in 2020. English contains fewer articles overall, with approximately 12,000–28,000 articles per year. These distributions demonstrate substantial coverage across all three languages while also highlighting differences in the volume and temporal availability of news content.

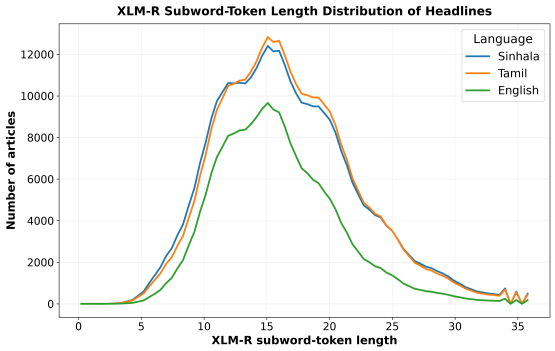


Figure 1: Token length distribution of headlines

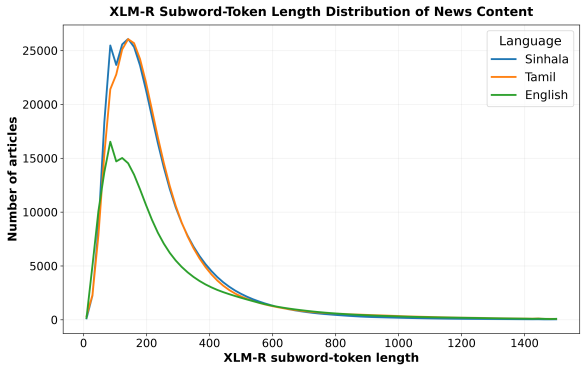


Figure 2: Token length distribution of contents

3 Benchmark Tasks

We derive three benchmark tasks using CEYNEWS that capture complementary aspects of news processing to show the utility of the corpus like CEYNEWS: (i) news source identification, (ii) news category classification, and (iii) news headline generation. The first two are classification tasks that assess whether models can capture outlet-specific and topic-specific patterns across Sinhala, Tamil, and English, while the third evaluates generative ability by requiring models to produce concise headlines from full articles. Together, these tasks provide a realistic testbed for analysing both language understanding and generation in the Sri Lankan multilingual context.

The task definitions, models, and evaluation metrics were informed by the experimental design of the NSINA corpus (Hettiarachchi et al., 2024), while being adapted to the trilingual setting and task requirements of CEYNEWS.

For each task, we constructed fixed training, validation, and test splits and evaluated the models under three experimental settings:

Language	Sentences	Tokens	Unique Types	Sent./Article	Tok./Article	Tok./Sent.	Median Tok./Sent.
Tamil	715,429	10.09 M	514,069	7.15	100.93	14.11	13
Sinhala	595,661	12.07 M	218,946	5.96	120.70	20.26	18
English	825,785	18.64 M	214,032	8.26	186.36	22.57	22

Table 2: Linguistic statistics computed from 100,000 randomly sampled articles per language using article content only. Tokens refer to word-level tokens.

1. **Supervised fine-tuning**: Models are fine-tuned on the combined trilingual training data and evaluated separately on each language. We focus on multilingual transformer models

2. **Few-shot evaluation**: LLaMA 3.1-8B-Instruct is evaluated without task-specific fine-tuning to assess few-shot performance on the defined tasks.

3. **Cross-lingual transfer**: Finally, the best-performing model from the supervised fine-tuning is trained on one language and evaluated on another to assess transferability between Sinhala and Tamil.

The following subsections describe the tasks, the machine learning models trained for each task, and the corresponding results.

3.1 Task 1: News Source Identification

News source identification classifies articles by their originating outlet using textual content. It may capture outlet-specific linguistic patterns, reflecting differences in editorial policy, tone, and lexical choice.

3.1.1 Data

For news source identification, the input consists of the cleaned article content, and the target label represents the originating news outlet. Source labels were automatically derived from the provenance metadata recorded during data collection rather than manually annotated. Articles collected from the Ada Derana, Hiru News, and ITN News websites were assigned the corresponding source labels. To prevent the models from relying on explicit source indicators, URLs, source attributions, reporter bylines, and website-navigation remnants were removed from the article content during preprocessing.

We randomly sampled 39,680 articles from each available language–source combination, resulting in a total of 317,440 articles across eight language–source combinations. ITN News was not included for English because it does not provide an English-language news section.

The data were divided into training, validation, and test sets using an 80/10/10 ratio with source stratified sampling. The data from all three languages were then combined, and the training set was randomly shuffled before model training.

3.1.2 Models & Experimental Setup

We evaluated four multilingual encoder models: XLM-R Base (Conneau et al., 2020), XLM-R Large (Conneau et al., 2020), InfoXLM Large (Chi et al., 2021), and mmBERT (Warner et al., 2025). All models were fine-tuned on the combined trilingual corpus for one epoch using a learning rate of 2×10^{-5}, a batch size of 16, and a maximum sequence length of 256. Based on trilingual fine-tuning performance, XLM-R Large was selected for cross-lingual transfer experiments, where it was trained on one language and evaluated on both in-language and cross-language test sets. We report Macro-F1 scores (Sokolova and Lapalme, 2009).

3.1.3 Results

Table 3 reports the Macro-F1 scores for news source identification. In the supervised fine-tuning setting, all multilingual encoders perform strongly, with XLM-R Large achieving the best results for Sinhala (0.967) and Tamil (0.812), and mmBERT performing best for English (0.949). These high scores across architectures suggest that news outlets exhibit strong outlet-specific structural and stylistic signatures – such as characteristic sentence length, punctuation conventions, formatting patterns, and recurring phrasing – that persist even after removing explicit source indicators during preprocessing, making the task highly learnable despite these controls.

Figure 3 shows confusion matrices for the best-performing model (XLM-R Large) across all three languages, providing insight into the lower Tamil performance noted above. For Sinhala, misclassifications are minimal and evenly rare across

all outlet pairs, consistent with its near-perfect Macro-F1. For Tamil, however, errors are concentrated asymmetrically: ITN News articles are frequently misclassified as Ada Derana (1,233 instances), substantially more than any other confusion pair, while Hiru News remains comparatively well-separated from both. This suggests that ITN and Ada Derana's Tamil-language editions may share more similar journalistic style or content conventions than either shares with Hiru. For English, which includes only two sources, confusion is moderate and roughly symmetric between the two classes.

In contrast, cross-lingual transfer is substantially weaker: models perform well on the language they are trained on but degrade sharply when evaluated on other languages, indicating limited transferability for source-specific stylistic features. Notably, transfer between the two local languages is consistently stronger than transfer to English in both directions. This asymmetry indicates that Sinhala and Tamil source-identification features transfer more readily between each other than either does to English.

Setting	Model	Sinhala	Tamil	English
Fine-tuning	InfoXLM Large	0.964	0.789	0.926
	mmBERT	0.943	0.802	**0.949**
	XLM-R Base	0.964	0.799	0.931
	XLM-R Large	**0.967**	**0.812**	0.935
Cross-lingual	Sinhala Fine-tuned	**0.968**	0.345	0.273
	Tamil Fine-tuned	0.390	**0.814**	0.114

Table 3: Macro-F1 scores for news source identification

3.2 News Category Classification

News category classification assigns articles to predefined topical categories based on their content. It is a standard NLP task with applications in information organisation and retrieval (Bracewell et al., 2009).

3.2.1 Data

For this task, the article content was used as the model input, and the news category served as the target label. Category labels were derived from the metadata provided by the original news websites rather than manually annotated. Ada Derana articles were excluded from this task because category information was not available on its individual article pages.

Hiru News provided five category labels: Local, International, Sports, Business, and Entertainment. ITN News used a broader and less consistent category scheme, including composite labels such as *Sports _ Local News* and *Breaking News _ International News*. We therefore applied a rule-based mapping to harmonise the source-specific labels into five common categories: Local, International, Sports, Business, and Entertainment. We then selected the four most frequent categories: (i) Local, (ii) International, (iii) Sports, and (iv) Business. Entertainment was excluded because it contained substantially fewer articles than the other categories.

For the Local, International, and Sports categories, we randomly selected 7,000 articles per language and category. For Business, all available Tamil and Sinhala articles were retained because of their limited availability. The resulting category-classification benchmark contains 81,309 articles in total.

The data were divided into training, validation, and test sets using an 80/10/10 ratio with category-stratified splitting. Data from all three languages were then combined, and the resulting multilingual training set was randomly shuffled before model training.

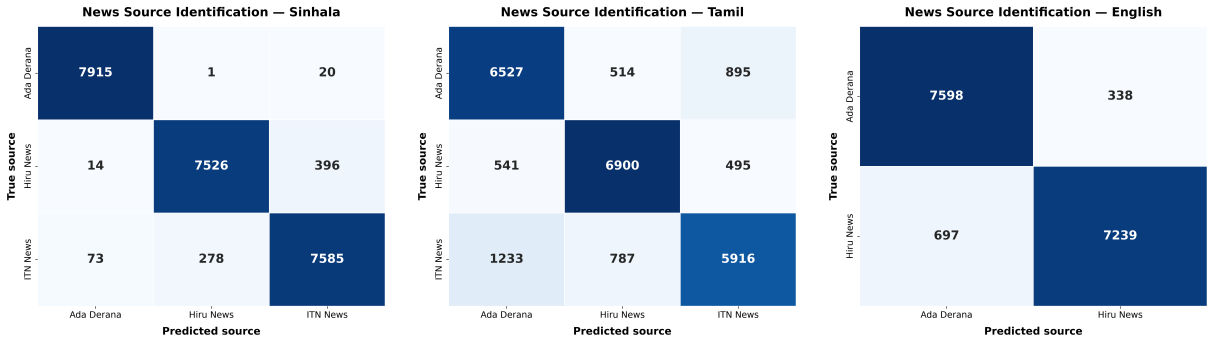
3.2.2 Models & Experimental Setup

We evaluated the same four multilingual encoder models used for source identification: XLM-R Base, XLM-R Large, mmBERT, and InfoXLM Large. All models were fine-tuned on the combined trilingual corpus for one epoch with a learning rate of 2×10^{-5}, a batch size of 16, and a maximum sequence length of 256. For cross-lingual transfer, we selected InfoXLM Large, the best-performing model in the trilingual fine-tuning setting.

3.2.3 Results

Table 4 reports the Macro-F1 scores for category classification. In the supervised fine-tuning setting, all models perform competitively across languages, with InfoXLM Large achieving the best scores for Sinhala (0.900) and Tamil (0.934), and mmBERT performing best for English (0.933).

Figure 4 shows confusion matrices for the best-performing model (InfoXLM Large) across all three languages. Sports remains almost perfectly separated in every language, consistent with its near-perfect recall. Local is the most error-prone true category across all languages, but its confusions are distributed differently in each: in Sinhala,



(a) Sinhala (b) Tamil (c) English

Figure 3: Confusion matrices for news source identification (XLM-R Large) across all three languages.

Local articles are most often misclassified as International, whereas in Tamil and English, Local errors are more evenly split between International and Business.

Cross-lingual transfer shows some degradation, but the drop is moderate compared to source identification, suggesting that category classification transfers better across languages. This is consistent with the per-class pattern above: Sports articles rely heavily on named entities and specialised terminology, including player names, team names, and tournament names that often remain recognisable even across scripts, whereas source identification depends on outlet-specific style, wording, and editorial conventions that are more language- and script-bound and therefore harder to transfer.

Setting	Model	Sinhala	Tamil	English
Fine-tuning	InfoXLM Large	**0.900**	**0.934**	0.930
	mmBERT	0.885	0.931	**0.933**
	XLM-R Base	0.894	0.928	0.928
	XLM-R Large	0.899	0.935	0.929
Cross-lingual	Sinhala	0.904	0.890	0.850
	Tamil	0.854	0.925	0.836

Table 4: Macro-F1 scores for news category classification

3.3 Headline Generation

Headline generation is a Natural Language Generation (NLG) task where the model generates a news headline from the full article content.

3.3.1 Data

For headline generation, the `News Content` field was used as the model input, and the corresponding `Headline` field was used as the target output. We extracted the headlines from the original news articles during data collection.

We used all available Tamil and English article–headline pairs, while the Sinhala dataset was down-sampled to approximately match the size of the Tamil dataset. The resulting benchmark contains 360,948 Sinhala articles (37.9%), 360,884 Tamil articles (37.9%), and 229,420 English articles (24.1%), giving a total of 951,252 article–headline pairs.

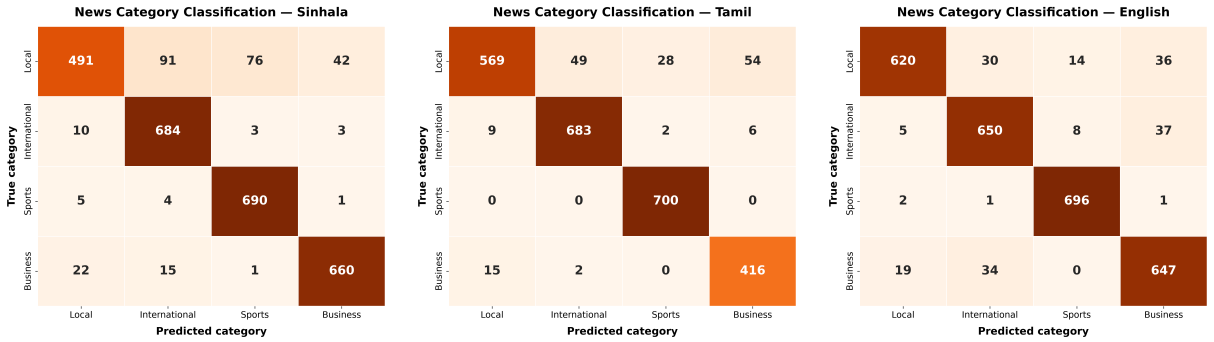
The data were divided into training, validation, and test sets using an 80/10/10 ratio. Data from all three languages were then combined, and the resulting multilingual training set was randomly shuffled before model training.

3.3.2 Models & Experimental Setup

We evaluated three multilingual sequence-to-sequence models: mBART-large-50 (Liu et al., 2020), mT5-base (Xue et al., 2021), and mT5-large (Xue et al., 2021). Models were fine-tuned on the combined trilingual corpus for one epoch with a learning rate of 1×10^{-5}, batch size of 8, maximum input length of 256, and maximum target length of 64. For cross-lingual transfer, we selected mT5-large. We report BLEU (Papineni et al., 2002) and Translation Edit Rate (TER) (Snover et al., 2006).

3.3.3 Results

Table 4 reports BLEU and TER scores for headline generation. In the fine-tuned setting, mBART-50 performs best for Sinhala and Tamil, achieving the highest BLEU and lowest TER, while mT5-large performs best for English and shows stable performance across all three languages. Compared to the classification tasks above, absolute BLEU scores are low across all languages and models (0.13–0.23) even in the fine-tuned setting. This is expected given the nature of the task: unlike classification, where each example has a single cor-



(a) Sinhala (b) Tamil (c) English

Figure 4: Confusion matrices for news category classification (InfoXLM Large) across all three languages.

rect label, headline generation is an open-ended summarisation task with many valid outputs for a given article, so n-gram overlap metrics such as BLEU and TER can penalise fluent, semantically appropriate headlines that do not closely match the exact reference wording. We report BLEU and TER as they are the standard metrics used in prior Sinhala headline generation work (Hettiarachchi et al., 2024). This limitation may be more pronounced for Sinhala and Tamil, whose morphologically rich, agglutinative structure allows a single word to take many valid inflected forms; exact lexical matching can therefore treat such variants as mismatches, which may contribute to the comparatively lower scores observed for these languages relative to English.

Cross-lingual transfer shows a substantial drop across all languages, with very low BLEU scores and high TER values. This indicates that headline generation is difficult to transfer across languages, likely because it requires not only content understanding but also language-specific generation and summarisation style or conventions. This is a stronger requirement than the pattern-matching underlying the classification tasks discussed earlier: whereas source identification and category classification rely on recognising structural or topical signals in text the model has already seen, headline generation requires producing fluent, grammatically well-formed output in a target language, which is substantially harder to achieve without in-language supervision.

3.4 Few-shot Evaluation

We evaluated a Llama model because its variants are openly available, relatively lightweight, and widely used in practical applications (Llama, 2023). Based on these considerations, we selected

		Sinhala		Tamil		English	
Setting	**Model**	**BLEU**	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
Fine-tuned	mBART-50	**0.167**	**0.946**	**0.234**	**0.811**	0.151	0.926
	mT5-base	0.134	1.058	0.201	0.877	0.183	0.815
	mT5-large	0.150	0.998	0.219	0.838	**0.205**	**0.779**
Cross-lingual	Sinhala	0.019	1.830	0.032	2.364	0.025	2.240
	Tamil	0.021	2.123	0.037	2.416	0.027	2.826

Table 5: BLEU and TER scores for headline generation

Llama 3.1-8B-Instruct for few–shot in–context learning evaluation. We did not use SinLlama (Aravinda et al., 2025) because it is a Sinhala-specific pretrained model and no instruction-tuned variant was available for a comparable prompting-based evaluation. For each task and language, we evaluated Llama on a class-stratified subsample of 500 instances drawn from the same held-out test split used for the fine-tuned models. This sampling strategy was used to make the evaluation computationally tractable while maintaining representation across classes.

For classification tasks, the number of demonstrations was set to the number of classes, so that each class was represented once in the prompt. Accordingly, we used 4-shot prompts for category classification, 3-shot prompts for Sinhala and Tamil source identification, and 2-shot prompts for English source identification. For headline generation, which does not involve discrete class labels, we used a fixed 3-shot prompt. All prompts were constructed using the Llama instruction chat template.

The detailed few–shot evaluation results for news source identification, news category classification, and headline generation, are provided in Tables 6, 7, and 8, respectively, in Appendix A. The model performs relatively better on English, achieving a Macro-F1 of 0.737 for category classification and 0.545 for source identification, while

obtaining a BLEU score of 0.067 for headline generation. In contrast, performance on Sinhala and Tamil remains extremely low across the evaluated tasks.

For Sinhala and Tamil, the model often generates corrupted or unreadable character sequences, suggesting tokenisation and script-handling issues.

4 Research Directions

The findings from this study point to several directions for future research. Building on CEYNEWS, future work can explore the following directions:

- **Improved cross-lingual transfer:** The weak transfer observed in source identification and headline generation suggests the need for better language alignment methods. Future work could explore bi-encoder architectures, multilingual contrastive learning, and shared representation alignment techniques to reduce the gap between in-language and cross-lingual performance.
- **Comparative multilingual news analysis:** Since CEYNEWS contains comparable news articles across Sinhala, Tamil, and English, it can support social science research on how the same or related events are reported differently across languages, such as framing, emphasis, sentiment, terminology, and narrative structure across language communities.
- **Parallel and semi-parallel corpus construction:** Although CEYNEWS is large-scale and trilingual, it is comparable rather than parallel. Future work could use cross-lingual news alignment techniques to identify article pairs or clusters describing the same events, enabling more controlled multilingual and translation-oriented research.
- **Script-robust multilingual models:** The poor few-shot performance of Llama 3.1-8B-Instruct on Sinhala and Tamil, including corrupted or unreadable outputs, suggests limitations in tokenisation and script representation. This also opens an important direction for future research.

5 Conclusion

This paper introduced CEYNEWS, a large-scale trilingual Sri Lankan news corpus comprising over 1.02 million articles and 136 million tokens in Sinhala, Tamil, and English, collected from three major outlets between 2013 and 2026. We derived three benchmarks – news source identification, news category classification, and headline generation – and evaluated multilingual models under supervised fine-tuning, cross-lingual transfer, and few-shot settings. The results show that fine-tuned multilingual models perform strongly, but cross-lingual transfer remains task-dependent: category classification transfers relatively well, while source identification degrades sharply across languages. Few-shot evaluation with LLaMA 3.1-8B-Instruct further reveals serious script-level robustness issues for Sinhala and Tamil. By releasing CEYNEWS, we aim to support multilingual NLP, cross-lingual transfer research, and more equitable language technology development for Sri Lankas multilingual communities.

6 Limitations

The main contribution of this work is the CEYNEWS corpus and its standardised multilingual benchmark splits. The reported model results are intended as baseline results to demonstrate the usefulness of the corpus, rather than as state-of-the-art results. To ensure a controlled comparison, all models were fine-tuned for one epoch using fixed hyperparameters across languages. More extensive hyperparameter tuning and ablation studies are left for future work.

7 Ethical Considerations

We considered ethical and privacy aspects during the construction and release of the CEYNEWS corpus. Before crawling, we examined the publicly available `robots.txt` files of all three news sources. The collection followed the available crawling directives and did not intentionally access paths explicitly disallowed by these files.

To reduce the risk of author profiling, the publicly released dataset has been randomized and does not preserve the original chronological or author-based ordering of the articles.

Acknowledgment

Ajintha Sivakulasingam and Kengatharaiyer Sarveswaran acknowledge the support received from Google through the Research Scholar Program and from the University of Jaffna through its research grant for carrying out this research.

References

HWK Aravinda, Rashad Sirajudeen, Samith Karunathilake, Nisansa de Silva, Rishemjit Kaur, and Surangika Ranathunga. 2025. Sinllama-a large language model for sinhala. In *2025 Moratuwa Engineering Research Conference (MERCon)*, pages 617–622. IEEE.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *Preprint*, arXiv:2110.06733.

David B. Bracewell, Jiajun Yan, Fuji Ren, and Shingo Kuroiwa. 2009. Category classification and topic discovery of japanese and english news articles. *Electronic Notes in Theoretical Computer Science*, 225:51–65. Proceedings of the Irish Conference on the Mathematical Foundations of Computer Science and Information Technology (MFCSIT 2006).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Anis Charfi, Wajdi Zaghouani, Syed Hassan Mehdi, and Esraa Mohamed. 2019. A fine-grained annotated multi-dialectal Arabic corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 198–204, Varna, Bulgaria. INCOMA Ltd.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. BERTifying Sinhala - a comprehensive analysis of pre-trained language models for Sinhala text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385, Marseille, France. European Language Resources Association.

Iakovos Evdaimon, Hadi Abdine, Christos Xypolopoulos, Stamatis Outsios, Michalis Vazirgiannis, and Giorgos Stamou. 2024. GreekBART: The first pretrained Greek sequence-to-sequence model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7949–7962, Torino, Italia. ELRA and ICCL.

Mohammad Ghaseminejad Raeini. 2025. The evolution of language models: From n-grams to llms, and beyond. *Natural Language Processing Journal*, 12:100168.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A news corpus for Sinhala. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.

Hansi Hettiarachchi, Tharindu Ranasinghe, Alistair Plum, Paul Rayson, Ruslan Mitkov, Mohamed Medhat Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage. 2026. Overview of the second workshop on language models for low-resource languages (LoResLM 2026). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 651–661, Rabat, Morocco. Association for Computational Linguistics.

Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Randunu Chandrakantha Uyangodage. 2025. Overview of the first workshop on language models for low-resource languages (LoResLM 2025). In *Proceedings of the First Workshop on Language Models for*

Low-Resource Languages, pages 1–8, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Varsha Jeyarajalingam, Menan Velayuthan, Sumirtha Karunakaran, Nivethiga Rasan, and Kengatharaiyer Sarveswaran. 2026. Evaluating linguistic knowledge of LLMs in Tamil: The ILAKKANAM benchmark. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC*, pages 176–189. European Language Resources Association (ELRA).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

R Kailainathan. 1999. Linguistic heritage of sri lanka tamil.

Bhadriraju Krishnamurti. 2003. *The dravidian languages*. Cambridge University Press.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Hugo Touvron Llama. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*.

Ahrane Mahaganapathy, Sumirtha Karunakaran, Kavitha Navakulan, and Kengatharaiyer Sarveswaran. 2026. Bridging dialectal variation: A phonetic transcription tool for Tamil. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 234–241, Rabat, Morocco. Association for Computational Linguistics.

Dushyanthi Mendis and Harshana Rambukwella. 2020. Sri lankan englishes. In *The Routledge handbook of world Englishes*, pages 178–193. Routledge.

Kurt Micallef and Claudia Borg. 2025. MELABenchv1: Benchmarking large language models against smaller fine-tuned models for low-resource Maltese NLP. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20505–20527, Vienna, Austria. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2025a. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*, 59(1):297–337.

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025b. MUSTS: MUltilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353, Vienna, Austria. Association for Computational Linguistics.

Tharindu Ranasinghe, Hansi Hettiarachchi, Nadeesha Chathurangi Naradde Vidana Pathirana, Damith Premasiri, Lasitha Uyangodage, Isuri Nanomi Arachchige, Alistair Plum, Paul Rayson, and Ruslan Mitkov. 2025c. Sinhala encoder-only language models and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8623–8636, Vienna, Austria. Association for Computational Linguistics.

Surangika Ranathunga, Asanka Ranasinghea, Janaka Shamala, Ayodya Dandeniyaa, Rashmi Galappaththia, and Malithi Samaraweeraa. 2025. A multi-way parallel named entity annotated corpus for english, tamil and sinhala. *Preprint*, arXiv:2412.02056.

Harold F Schiffman. 1999. *A reference grammar of spoken Tamil*. Cambridge University Press.

Nuwan I. Senaratna. 2026. Sri lanka document datasets: A large-scale, multilingual resource for law, news, and policy. *Preprint*, arXiv:2510.04124.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 2 others. 2024a. An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 38–46, Torino, Italia. ELRA and ICCL.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François,

Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024b. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Adsajan Thillainathan, Nishanthini Kanthakumar, Nivethiga Rasan, and Kengatharaiyer Sarveswaran. 2026. TaLK-Corpus: A Regionally Diverse Evaluation Set for Sri Lankan Tamil Speech. In *Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) at LREC2026*, pages 194–200. European Language Resources Association (ELRA).

Naveed Unjum, Stephanie Evert, Kengatharaiyer Sarveswaran, Ruvan Weerasinghe, and Nevidu Jayatilleke. 2025. Llms for low-resource languages: A survey. *Available at SSRN 6122026*.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

A Few-shot Evaluation Results

This appendix provides the detailed results of the few–shot evaluation described in Section 3.4. The evaluation covers three tasks: headline generation, news category classification, and news source identification. Results are reported separately for Tamil, Sinhala, and English using the same evaluation settings described in the main paper. For headline generation, BLEU and Translation Edit Rate (TER) scores are reported, while Macro-F1 is reported for the two classification tasks.

Language	Macro-F1
Sinhala	0.0040
Tamil	0.0040
English	0.5448

Table 6: News source identification results in the few-shot setting.

Language	Macro-F1
Sinhala	0.0000
Tamil	0.0000
English	0.7366

Table 7: News category classification results in the few-shot setting.

Language	BLEU	TER
Sinhala	0.0004	7.2935
Tamil	0.0013	5.5479
English	0.0667	10.5605

Table 8: Headline generation results in the few-shot setting.

The TER scores should be interpreted together with the generated headline length and quality. The English model generally produces longer and more complete headlines, which can result in a higher number of edits when the generated headline differs in length from the reference. Although the generated headline is semantically related, its different wording and length lead to a higher TER. In contrast, the Sinhala and Tamil outputs are often very short or incomplete. Therefore, TER alone does not fully reflect the quality of the generated headlines.

Probing Pretraining "Quality": Sociolinguistic Mobility as a Lens on Web Corpus Filters

Jessica C. Wang
University of Washington
Department of Linguistics
jcmw614@uw.edu

Abstract

Data selection algorithms for large language model (LLM) pretraining are designed to mine 'quality' from the web, but that notion is murky at best. This study explores whether sociolinguistic mobility – realized through language use, in context – offers a lens for understanding what such algorithms actually select. As a case study, the FineWeb pretraining pipeline is examined for how it reshapes two tangible axes of language use: the geographical provenance and register composition of texts drawn from ICE and GloWbE. FineWeb rejection rates are highest, and fairly uniform across provenance groups, for web-native and high-context registers (90.9% and 99.2%, respectively) but within medium-context registers, association with the United States is moderately protective: the 7% rejection rate is well below the medium-context average of 38%, an effect corroborated by NPMI. Unlike prior work on learned, classifier-based filters and US-internal sociolect bias, this study finds comparable effects in a deterministic pipeline within the global spread of English. As pretraining data selection becomes increasingly opaque, these results suggest that algorithmic 'quality' does not represent a real-world range of English use manifested in the web-as-corpus – a gap made visible by tangible sociolinguistic dimensions.

1 Introduction

The construction of LLMs today relies on massive web-scraped archives, popularly perceived as 'raw' data from which objectively valuable knowledge can be extracted, despite growing critical scholarship that complicates this framing (Baack, 2024; Desai et al., 2024). This epistemological assumption has motivated the deployment of various data curation protocols to sanitize LLM pretraining input, including a range of 'quality' filters (Albalak et al., 2024). Indeed, the world-wide web hosts myriad linguistic material (Crystal, 2011; Blommaert, 2018; Mair, 2025) of varying utility for NLP tasks – and yet, the manner in which data curation manipulates the "maturing medium" (Tinnefeld, 2023) of internet language has eluded full capture. Sociolinguistic perspectives on language modeling provide insight into this open area of research: "the validity and value of a language model depends on the degree to which the training corpus accurately represents the variety that is effectively being modeled [...] even if that variety of language is unknown or under-specified" (Grieve et al., 2025, p. 5). This raises the question of which dimensions of linguistic diversity matter most as LLM-based technologies are exposed to an increasing number of social and communicative contexts.

To that end, this work begins by presenting a paradigm of sociolinguistic *mobility*, following Blommaert (2010), as a lens to describe the variation in English, manifested on the web-as-corpus (§2). This motivates the central study, wherein the FineWeb quality filtering pipeline (Penedo et al., 2024) is applied to a corpus compiled from ICE and GloWbE, spanning a range of geographic and register contexts (§3). The results show that register context level is strongly associated with filter rejection, with a further interaction between register and geographic provenance at the medium-context level (§4), suggesting that sociolinguistic axes can help describe algorithmic behavior. §5 discusses directions for future work, including more granular analysis of the low-context subset, where rejection patterns are more volatile and provenance-level trends remain preliminary. Experimental code will be made available on GitHub [1].

2 Sociolinguistic Mobility and Web-as-Corpus

The prevailing paradigm for linguistic variation centers on the *distribution* of language varieties in

[1] https://github.com/jessiclassy/pretraining-quality-mobility

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 320–330
October 29, 2026 ©2026 Association for Computational Linguistics

place and time (Kohnen and Mair, 2012; Trudgill, 2017; Kortmann et al., 2020; Szmrecsanyi and Grafmiller, 2023a), characterized by a range of features (e.g., phonological, lexical and/or morphosyntactic). In effect, this paradigm treats varieties as fixed, discrete classes to be identified and balanced across a dataset, as if representation were a classic machine learning problem. This framing has been widely embraced by contemporary dialect-centered NLP approaches (Joshi et al., 2025) and continues to inform directions for language technology development (Bird, 2022), such as language-variety benchmarks built via synthetic data generation, simulating regional dialects as 'translations' of standard English (Ziems et al., 2022, 2023; Lee et al., 2025; Oh et al., 2026). But treating varieties as classes to be balanced takes them as a given, when they have evolved (and continue to develop) from ongoing geographic and sociohistorical processes: contact, migration, standardization.

A more recent paradigm, emerging from works such as Pennycook (2006) and Blommaert (2010), reframes langauge varieties as one among several tangible resources, including register and genre, which are continually produced and reused as speakers move through the social, spatial and temporal trajectories of their lives. Blommaert makes the stakes of this shift explicit:

> The crux of the matter is that we need to think of issues such as linguistic inequality as being organized around concrete resources, not around languages in general but specific registers, varieties, genres. And such concrete resources follow the predicament of their users: when the latter are socially mobile, their resources will follow this trajectory; when they are socially marginal, their resources will also be disqualified. [...] If we fail to do that, we risk drawing a caricature of social realities and becoming very upset about that caricature rather than about an accurate replica of social processes. (Blommaert, 2010, p. 47)

The distributional paradigm's flaw, in this view, is not that it attends to variety, but that it treats variation as a static inventory rather than the product of ongoing sociohistorical processes. This inadvertently ignores how a linguistic resource may be disqualified: users who are socially marginal have their resources disqualified alongside them, a concern raised over a decade ago and still echoed by sociolinguists writing on LLMs (Grieve et al., 2025; Nguyen, 2025; Lutgen et al., 2026).

Provenance group	Regions
1 (Traditional L1)	United States, Great Britain
2 (High-contact L1)	Australia, Canada, Ireland, New Zealand
3 (Indigenized L2)	Singapore, Hong Kong, India, Philippines,
4 (High-contact L2)	Jamaica, Kenya, Nigeria, Tanzania

Table 1: Geographic provenance, grouped by sociohistorical patterns of English adoption.

Pretraining data filters risk enacting a comparable disqualification: each filtering decision determines which linguistic resources move forward into a corpus, a model, and its downstream outputs – and which do not. Capturing this process requires attention to at least two things distribution leaves flat: the sociohistorical conditions from which a resource emerges, and the situational context in which it is deployed. This study operationalizes the former as geographic provenance and the latter as register context level, examined jointly.

English is a strategic case for this kind of analysis: no other language is as widely used on the web, nor carries as much internal variation as a consequence of its complex history of spread (Kortmann et al., 2020; Pimienta, 2022). This makes it an unusually rich site for observing how sociolinguistic resources are manipulated in the data that trains today's langauge models.

3 Probing Pretraining Filters

To probe at the notion of pretraining 'quality' in terms of geographic provenance and register context level, this study applies the open-source FineWeb pretraining filter pipeline (Penedo et al., 2024) to texts drawn from existing English corpora. Resulting document rejection rates are analyzed across both dimensions of sociolinguistic variation as a practical measure of web corpus filter behavior, supplemented by normalized pointwise mutual information.

3.1 Representing Sociolinguistic Mobility

3.1.1 Existing corpora

Data from fourteen nations with attested development of local English dialects are drawn from two

Context level	Examples
low	printed academic writing, legal presentations
medium	personal correspondence, scripted conversations, press editorials, instructional/creative writing
high	private/unscripted conversations and monologues

Table 2: Register context levels, grouped by situational characteristics, such as participation, channel, production circumstances, and setting.

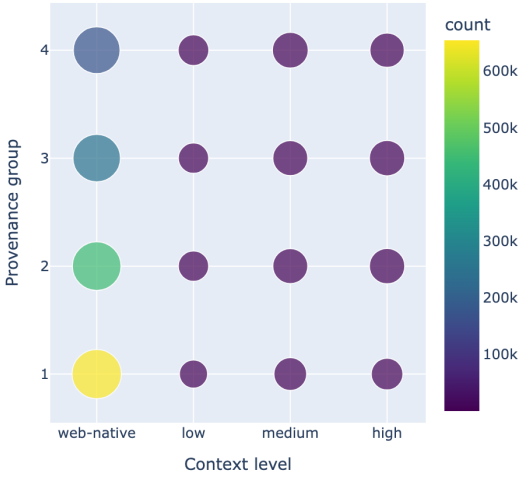


Figure 1: Data sizes with respect to provenance group and register context level.

existing corpora: the Corpus of Global Web-based English (Davies and Fuchs, 2015), abbreviated as GloWbE, and the International Corpus of English (Greenbaum, 1991; Kirk and Nelson, 2018), abbreviated as ICE. GloWbE was collected across a six-month period of global usage of web-native English, scraping over 1.5M web pages from Google's search index. ICE, by contrast, consists entirely of non-web documents, but is uniquely detailed in its geographical and situational annotation. Over the past two decades and across more than a dozen academic institutions, teams of linguists have manually curated each regional component to capture approximately one million words of prototypical, local English, targeting 60% transcribed speech and 40% written texts. All regional components follow a shared annotation schema to flag editorial information (e.g., overlapping speech, normative corrections), as well as a document genre typing (e.g., scripted or unscripted speech, academic writing, editorials, correspondence). We posit that this representation makes an important proxy for web text types such as institutional/professional texts, podcast or broadcast transcripts, or even the disfluent, multi-turn language found in web forums, chat logs or comment sections. Both existing corpora are considered essential resources to large-scale and fine-grained research on English variation (Mair, 2017; Szmrecsanyi and Grafmiller, 2023b; Qin, 2023; Collins, 2024; Basile, 2025). Thus, the dataset resulting from combining both corpora enables a web-scale analysis of how filter behavior is affected by geographical provenance, as well as a small-scale analysis of register context level, both independently and in combination with provenance.

3.1.2 Preprocessing

To align these documents with FineWeb's expected plaintext format, a registry of regular expressions is compiled to strip $> 99.9\%$ of ICE-specific intra-document feature annotations so that only the naturalistic content remains. No punctuation marks are stripped or added to ICE transcribed speech. Similarly, residual HTML tags and fair-use masking strings are removed from GloWbE documents (see Appendix C). Finally, both corpora are adapted to JSONL format.

To enable cross-corpus analysis of geographical provenance, all documents are grouped as described in Table 1, derived from the sociohistorical patterns in which English was adopted in a specific region (Trudgill, 2017; Kortmann et al., 2020). ICE-specific register labels are also categorized into three coarse situational-context categories: low, medium, high, as shown in Table 2. These categories are assigned through a qualitative grouping of ICE genre labels by the level of situational context needed between participants, guided by four of the characteristics identified in Biber and Conrad (2009)'s descriptive framework of situational register: participants, channel, production circumstances, and setting. In the absence of curated register labels for GloWbE documents, they are analyzed as a whole under a general 'web-native' category whose situational context is underdetermined.

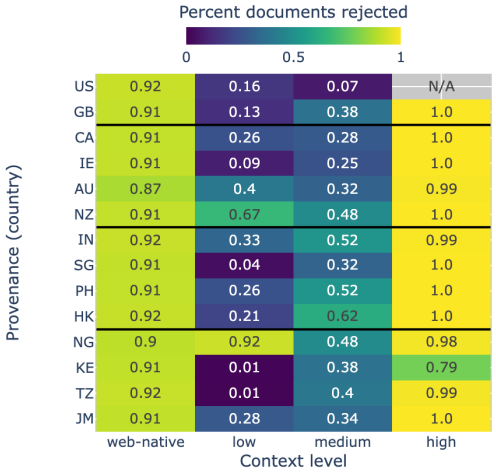


Figure 2: Heatmap of reject rates, aggregated by geographical provenance and register context levels. Solid lines indicate provenance groups.

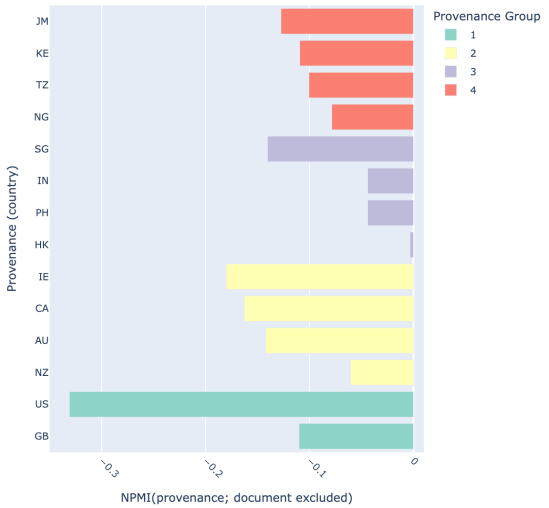


Figure 3: Normalized pointwise mutual information score for geographical provenance and filter rejection, within the medium register context level.

3.2 Analyzing the FineWeb Quality Pipeline

For this experiment, the combined English corpus is processed through the five stages of FineWeb pipeline (Penedo et al., 2024) to obtain per-document rejection decisions, which are aggregated over the two sociolinguistic axes for analysis. These stages cover English language identification, repetition filters, and several deterministic quality filters which screen documents for specific character/word/line thresholds. Filtering algorithm details can be found in Appendix B.

4 Results & Discussion

Figure 2 presents a heatmap of document rejection rates, aggregated by geographical provenance and register context levels for all fourteen countries represented in both GloWbE and ICE. Broadly speaking, all provenance groups follow the same trend: web-native GloWbE subsets pattern with high-context texts, both showing the highest overall rejection rates, averaging at 90.9% and 99.2% respectively. Within medium-context texts, association with the United States appears to have a moderately protective effect, with a 7.1% rejection rate compared to the average medium-register rate of 38.4%. This is corroborated by normalized pointwise mutual information (NPMI $\in [-1, 1]$, where 0 indicates independence and negative values indicate avoidance), shown in Figure 3 (NPMI=-0.33). The FineWeb pipeline largely discards resources of primary importance to corpus linguistics and sociolinguistic traditions, reproducing the dominance of mainstream, US-centered linguistic contexts. This pattern emerges from an algorithmic mechanism distinct from the classifier-based biases identified in prior work (Dodge et al., 2021; Gururangan et al., 2022; Lucy et al., 2024; Deas et al., 2025). In other words, the notion of 'quality' remains divorced from a real-world range of language use.

Rejection rates for low-context texts are much more volatile, and are not analyzed in depth here. More than half of the geographies represented among low-context texts are more protected from filter rejection than their medium-context counterparts by at least 10%, though patterns vary considerably by country: association with the United States, Canada, Australia, and Jamaica yield similar rejection rates to their medium-context counterparts, while association with Nigeria and New Zealand yield far greater ones. Given the small size of this low-context subset ($11 \leq n \leq 69$ for each country), these patterns are best treated as preliminary rather than conclusive, with a more granular analysis of low-context text left to future work.

5 Conclusion

This study finds that provenance and register, examined jointly, reveal structure in FineWeb's filtering behavior that neither dimension shows on its

own: rejection is strongly associated with register context. Within medium-context text specifically, association with the United States is moderately protective against rejection, an effect that attending to language variety alone, without the sociohistorical conditions of its production or situational use, would not have surfaced this interaction. As the practice of LLM pretraining data curation continues to grapple with what "quality" text means, articulating the dimensions of web-as-corpus through sociolinguistic terminology offers a concrete vocabulary, capable of describing not just what a filter selects, but what kind of language use it renders legible or illegible in the process.

Beyond the filtering pipeline studied here, this approach could be extended to compare pretraining data curation strategies across filters (Lucy et al., 2024; Deas et al., 2025), or applied to data mixing protocols (Chen et al., 2026), where sociolinguistic sensitivity may operate differently than under binary retention decisions. More broadly, it appears that a resource-oriented, mobility-based perspective may provide an initial step toward a fuller diagnostic vocabulary for evaluation what data-centric curation choices encode about language on the web.

Limitations

Simplicity of the mobility framework. The mobility framework operationalized here reduces a rich theoretical construct to a handful of discrete levels, inferred from corpus provenance and genre/register conventions. This is a deliberately coarse instrument, sufficient to establish whether filtering outcomes are sensitive to mobility at all, but not to capture finer-grained distinctions within or across mobility levels. Future work might operationalize mobility more directly, for instance through document-level features rather than corpus-level proxies.

Parity of ICE and GloWbE. The two corpora unified for this study were built using very different methods, capturing English variation across differing communicative contexts, and resulting in starkly different data volumes. While the present study treats ICE and GloWbE jointly to explore the diversity represented among these texts, any direct comparison of results between the two corpora should be treated with caution, partly because GloWbE lacks register/genre annotation comparable to ICE's. Future work may leverage automatic register classification labels (Laippala et al., 2023) to expand upon the ICE-centric results in the present study.

Statistical rigor of rejection patterns. The present study focuses concrete measurements of rejection rates across the dimensions of geographic provenance and register context levels, supported by normalized pointwise mutual information scores to describe the site of interaction between these variables. This appraoch is especially limited for sparse subsets, such as low-context text, where country-level sample sizes are small (as few as 11 documents) and rejection rates are correspondingly volatile. The interpretation of such results is limited, but points to promising avenues which could be more rigorously examined with mixed effect modeling, as is common in the sociolinguistic tradition (Gudmestad and Metzger, 2025, *inter alia*), or perplexity analyses (Deas et al., 2025; Dunn, 2024).

Treating FineWeb as a black box. This work is oriented towards a single pretraining filtering pipeline as a practical approximation of how filtering preferences are manifested, but in doing so, the pipeline itself is treated as a black box. Such filtering pipelines actually consist of a sequence of deterministic text filters. Detailed characterization of the rejection patterns specific to each stage or specific text heuristic are out of scope for the present study, but the progression of rejection patterns across pipeline stages is available in Appendix B, and an initial breakdown of text heuristic patterns are available in Appendix D.

Acknowledgments

This paper benefited from helpful conversations with Florian Humbert and thoughtful guidance from Amandalynne Paullada.

Parts of this work was completed on Hyak, UW's high performance computing cluster. This resource was funded by the UW student technology fee.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research*.

Stefan Baack. 2024. A critical analysis of the largest source for generative ai training data: Common crawl.

In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2199–2208, New York, NY, USA. Association for Computing Machinery.

Carmelo Alessandro Basile. 2025. Modal better: A corpus-based investigation in world englishes. *World Englishes*, 44(3):318–338.

Douglas Biber and Susan Conrad. 2009. *Describing the situational characteristics of registers and genres*. Cambridge Textbooks in Linguistics. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Jann Blommaert. 2010. *The Sociolinguistics of Globalization*. Cambridge Approaches to Language Contact. Cambridge University Press.

Jann Blommaert. 2018. Durkheim and the internet.

Mayee F. Chen, Tyler Murray, David Heineman, Matt Jordan, Hannaneh Hajishirzi, Christopher Ré, Luca Soldaini, and Kyle Lo. 2026. Olmix: A framework for data mixing throughout lm development. *Preprint*, arXiv:2602.12237.

Peter Collins. 2024. Variation in world englishes through the lens of negation. *World Englishes*, 43(1):47–70.

David Crystal. 2011. *Internet Linguistics: A Student Guide*. Taylor Francis Group, Oxford, UNITED KINGDOM.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus (glowbe). *English World-Wide. A Journal of Varieties of English*, 36(1):1–28.

Nicholas Deas, Blake Vente, Amith Ananthram, Jessica Grieser, Desmond Patton, Shana Kleiner, James Shepard, and Kathleen McKeown. 2025. Data caricatures: On the representation of african american language in pretraining corpora. *CoRR*, abs/2503.10789.

Meera A. Desai, Irene V. Pasquetto, Abigail Z. Jacobs, and Dallas Card. 2024. An archival perspective on pretraining data. *Patterns (New York, N.Y.)*, 5(4):100966.

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*, page 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Dunn. 2024. Pre-trained language models represent some geographic populations better than others.

Sidney Greenbaum. 1991. Ice: the international corpus of english. *English Today*, 7(4):3–7.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7.

Aarnes Gudmestad and Thomas A. Metzger. 2025. Mixed-effects modeling with a multinomial dependent variable. *Language Learning*, 75(1):212–249.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6):149:1–149:37.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. (arXiv:1612.03651). ArXiv:1612.03651 [cs].

John Kirk and Gerald Nelson. 2018. The international corpus of english project: A progress report. *World Englishes*, 37(4):697–716.

Thomas Kohnen and Christian Mair. 2012. *Technologies of communication*, 1 edition, page 261–284. Oxford University Press.

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *The Electronic World Atlas of Varieties of English*. Zenodo.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.

Jiyoung Lee, Seungho Kim, Jieun Han, Jun-Min Lee, Kitaek Kim, Alice Oh, and Edward Choi. 2025. Trans-env: A framework for evaluating the linguistic robustness of llms against english varieties. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. Aboutme: Using self-descriptions in webpages to document the effects of english pretraining data filters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7393–7420, Bangkok, Thailand. Association for Computational Linguistics.

Anne-Marie Lutgen, Alistair Plum, Verena Blaschke, Barbara Plank, and Christoph Purschke. 2026. Variation is the norm: Embracing sociolinguistics in nlp. page 11332–11344.

Christian Mair. 2017. *World Englishes and Corpora*, page 0. Oxford University Press.

Christian Mair. 2025. The digital diffusion of pidgin/creoles: Langwij, kolcha, media and large language models. *Journal of Pidgin and Creole languages*.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

Jio Oh, Paul Vicinanza, Thomas Butler, Steven Euijong Whang, Dezhi Hong, and Amani Namboori. 2026. Should llms, *like*, generate how users talk? building dialect-accurate dialog[ue]s beyond the american default with mdial.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. (arXiv:2406.17557). ArXiv:2406.17557 [cs.CL].

Alastair Pennycook. 2006. *Global Englishes and Transcultural Flows*. Routledge, London.

Daniel Pimienta. 2022. Resource: Indicators on the presence of languages in internet. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, page 83–91, Marseille, France. European Language Resources Association.

Melissa Xiaohui Qin. 2023. A corpus-based approach to chinese english lexis. *World Englishes*, 42(2):308–326.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2021. Scaling language models: Methods, analysis amp; insights from training gopher.

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.

Benedikt Szmrecsanyi and Jason Grafmiller. 2023a. *Comparative Variation Analysis: Grammatical Alternations in World Englishes*, 1 edition. Cambridge University Press.

Benedikt Szmrecsanyi and Jason Grafmiller. 2023b. *The Data*, page 56–81. Studies in Language Variation and Change. Cambridge University Press, Cambridge.

Thomas Tinnefeld. 2023. Interview: 10 questions for professor david crystal. *Journal of Linguistics and Language Teaching*, 14(2):203–239.

Peter Trudgill. 2017. The spread of english. In Markku Filppula, Juhani Klemola, and Devyani Sharma, editors, *The Oxford Handbook of World Englishes*, page 0. Oxford University Press.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? (arXiv:1905.07830). ArXiv:1905.07830 [cs].

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Source Corpora Metadata

Specific document counts for GloWbE and ICE are presented in Tables 3 and 4.

B Filtering Pipeline

The FineWeb filtering pipeline (Penedo et al., 2024) consists of five stages, described below. Data volumes after every stage are presented in Figure 4, and shows that the most aggressive filter is in fact the FineWeb quality filter.

Country	Abbreviation	Web pages
United States	US	275,156
Canada	CA	135,692
Great Britain	GB	381,841
Ireland	IE	102,147
Australia	AU	129,244
New Zealand	NZ	82,679
India	IN	113,765
Singapore	SG	45,459
Philippines	PH	46,342
Hong Kong	HK	43,936
Nigeria	NG	37,285
Kenya	KE	45,962
Tanzania	TZ	41,356
Jamaica	JM	46,748
TOTAL		1,527,612

Table 3: GloWbE metadata

Country	Abbreviation	Written	Spoken	Total
United States	US	200	0	200
Great Britain	GB	200	300	500
Canada	CA	200	300	500
Ireland	IE	200	300	500
Australia	AU	199	300	499
New Zealand	NZ	200	300	500
India	IN	200	300	500
Singapore	SG	200	300	500
Philippines	PH	200	300	500
Hong Kong	HK	200	300	500
Nigeria	NG	504	392	896
Kenya	KE	181	167	348
Tanzania	TZ	175	115	290
Jamaica	JM	200	299	499
TOTAL		3,059	3673	6,732

Table 4: ICE metadata

B.1 Language Identification

This algorithmic filter will exclude documents that do not produce a language identification score $p(\text{eng}) > 0.65$. The language identification model is the `fasttext` classifier (Joulin et al., 2016).

B.2 Gopher Repetition Filter

This algorithmic filter computes the fraction of duplicate lines, paragraphs, and character-level n-grams, and excludes documents when any of those fractions exceed a hardcoded threshold. It was originally deployed to pretrain the Gopher model (Rae et al., 2021).

B.3 Gopher Quality Filter

This algorithmic filter also originates from (Rae et al., 2021), but is defined instead by applying a variety of text heuristics: number of words, mean word length, symbol-to-word ratio, bullet point presence, and stop word presence.



Figure 4: Data volumes after every stage; Stage 0 represents the initial volume before any filtering.

B.4 C4 Quality Filter

This algorithmic filter was deployed to create the prominent pretraining dataset C4 (Raffel et al., 2020), and excludes documents by applying another set of text heuristics: presence of the words "Javascript", "lorem ipsum", "terms of service", or curly brackets. Additionally, all lines without terminal punctuation are penalized such that if a document is found to have fewer than 5 non-penalized sentences or contain non-penalized lines with 3 or fewer words, it is discarded.

B.5 FineWeb Quality Filter

This algorithmic filter was designed specifically for the FineWeb dataset (Penedo et al., 2024), designed to be applied after the preceding filters and maximize aggregate accuracy on several prominent NLU benchmarks, such as HellaSwag (Zellers et al., 2019). In actuality, its definition is based entirely off of deterministic line-punctuation, newline, and character duplicate ratios.

C Source Corpus Examples

For brevity, the source corpus examples below are truncated to the first 1000 characters.

C.1 GloWbE Sample

C.1.1 Cleaned Text

About this item
As stated in the thesis project , " My community consists of over 650 non-profit organizations .These agencies range in size from the small local grassroots organization to fully staffed , fully funded government/private agencies .They consist of Soup Kitchens , Daycares and Senior Centers , M/R and M/H facilities , Drug/alcohol rehabs and Churches of all different denominations .The common thread that these agencies hold is the fact that they are all " social service agencies " that contain a feeding program that serves low-income individuals in some form whether through hot-prepared on-site meals or emergency food packages .The other common thread that they all share is membership into the Greater Philadelphia Food Bank which is a food collection , storage and distribution center . The collective mission of my community is to end " Hunger " .The need as identified through my observation as Agency Relations Director is to strengthen and focus these social service ag...

C.1.2 Original Text

<h> About this item <p> As stated in the thesis project , " My community consists of over 650 non-profit organizations . These agencies range in size from the small local grassroots organization to fully staffed , fully funded government/ private agencies . They consist of Soup Kitchens , Daycares and Senior Centers , M/R and M/H facilities , Drug/alcohol rehabs and Churches of all different denominations . The common thread that these agencies hold is the fact that they are all " social service agencies " that contain a feeding program that serves low-income individuals in some form whether through hot-prepared on-site meals or emergency food packages . The other common thread that they all share is membership into the Greater Philadelphia Food Bank which is a food collection , storage and distribution center . The collective mission of my community is to end " Hunger " . Each individual organization has its own defined method of achieving that result in conjunction with other goals ..

C.2 ICE Sample

C.2.1 Cleaned Text

Chandler and Frasier: Kampala's Most Wanted | Ernest Bazanye

"Well done."

"What?"

The old woman continued to smile broadly, all teeth and good nature. Her eyes were so squinted up with pleasant spirit that only bits of the black parts were visible. "Well done," she repeated, confident that the young man could understand her.

The young man could not. Chandler was a thoroughly and totally and completely worthless schoolboy so he was not used at all to hearing those words, together, in that order. He met the repetition of her statement with the repetition of his own. "What?"

This is when his elder brother Frasier joined in the conversation, brusquely nudging in and scowling. "She means 'Mugyebale' , you doofus

```
    ."

 Well done was, after all, the exact
     transliteration of the common
     Luganda greeting Mugyebale and,
     this was Kampala, the city where
     most people spoke first and worried
      only later, if ever, whether their
      speech had made any sense. It was
     also a common thing for tee...
```

C.2.2 Original Text

```
     <I> <h> <ICE-UG:W2F-016#1:1>
         Chandler and Frasier: <w>
         Kampala's </w> Most Wanted |
         Ernest Bazanye </h>

<p> <quote> <ICE-UG:W2F-016#2:1> "Well
    done." </quote> </p>

<p> <quote> <ICE-UG:W2F-016#3:1> "What?"
     </quote> </p>

<p> <ICE-UG:W2F-016#4:1> The old woman
    continued to smile broadly, all
    teeth and good nature. <ICE-UG:W2F
    -016#5:1> Her eyes were so squinted
    up with pleasant spirit that only
    bits of the black parts were visible.
     <quote> <ICE-UG:W2F-016#6:1> "Well
    done," </quote> she repeated,
    confident that the young man could
    understand her. </p>

<p> <ICE-UG:W2F-016#7:1> The young man
    could not. <ICE-UG:W2F-016#8:1>
    Chandler was a thoroughly and
    totally and completely worthless
    schoolboy so he was not used at all
    to hearing those words, together, in
     that order. <ICE-UG:W2F-016#9:1> He
     met the repetition of her statement
     with the repetition of his own. <
    quote> <ICE-UG:W2F-016#10:1> "What?"
     </quote> </p>

<p> <ICE-UG:W2F-016#11:1> This is when
    his elder brother Frasier joined in
    the...
```

D Filter Reason Breakdown

In Figure 5, The high-context facet plot shows 0% rejection for the US column because there is no high-context US data available. In contrast, the low-context facet plot shows 0% rejection for Kenya and Tanzania because none of the 22 legal presentation documents were rejected by FineWeb.

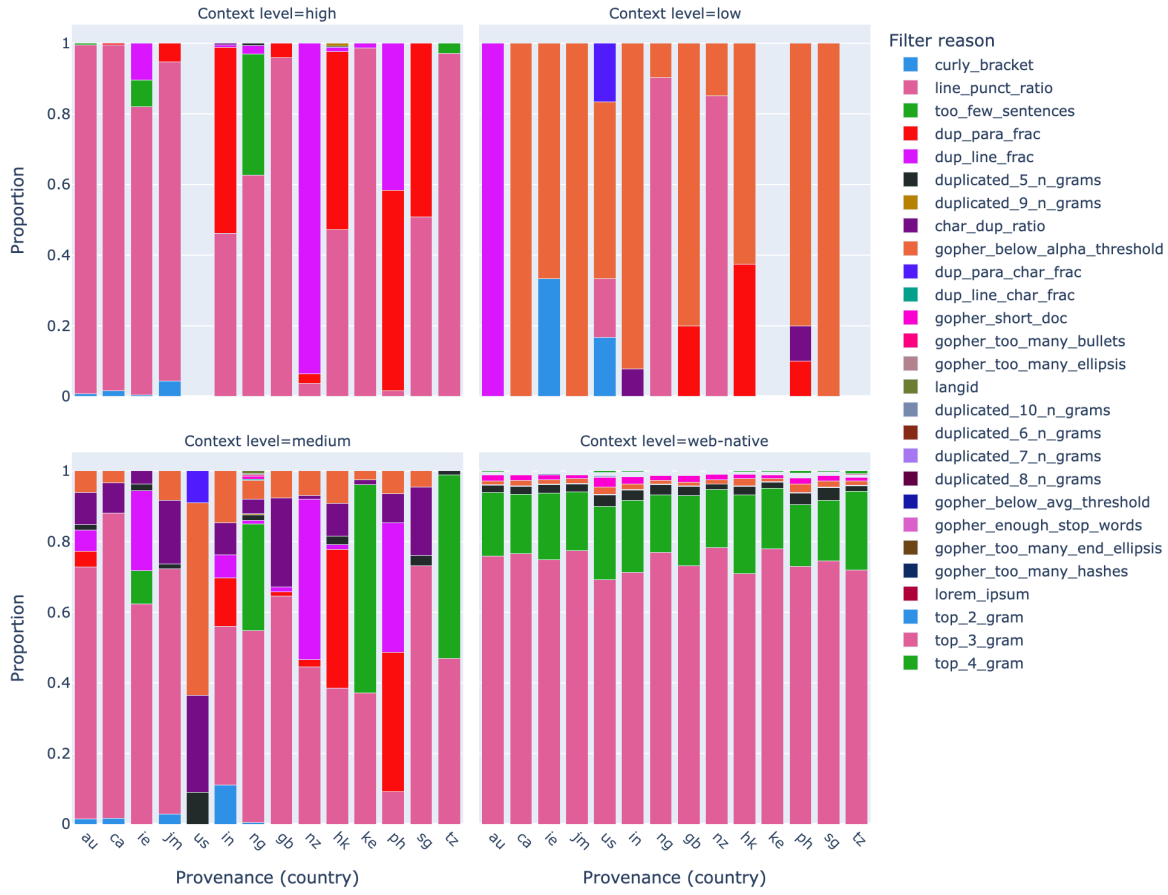


Figure 5: The relative proportions of filter rationales, aggregated by geographic provenance and register context levels.

A web-derived typology of multimodal infant crying: Taxonomy, a text-only web-corpus comparison, and ethical issues

Jacob Rigal
Université Paris Cité / LLF, CNRS
jacob.rigal@u-paris.fr

Jonathan Ginzburg
Université Paris Cité / LLF, CNRS
yonatan.ginzburg@u-paris.fr

Abstract

Web video platforms contain naturally occurring audiovisual interaction, but their use as corpora requires explicit decisions about selection, metadata, annotation, privacy, and redistribution. We present a web-derived data set of infant cry contexts coding 250 publicly accessible videos from YouTube, Instagram, and TikTok, covering vaccination, discomfort, gas, fussiness, colic-like crying, hunger, birth, sleepiness, overstimulation, and other pain contexts. Candidate videos were labeled from audiovisual evidence; validation currently covers 230 videos (92.0%), of which 226 receive a current taxonomy label and enter the audiovisual distributional tables. Validation also served to refine the taxonomy, revealing systematic under-recognition of sleep-related visual cues and recurrent subtypes within the broader *Fussiness* category. To test what can be recovered from linguistic material alone, we contextually classified 4,600 *baby + CRY* concordance lines from enTenTen21 with OpenAI GPT-5.6 Sol, recording a category, confidence level, and explanation for each line. A human audit of the first 460 lines (10.0%) covered all output categories and yielded 91.5% normalized categorical agreement ($\kappa = .885$; bootstrap 95% CI [.849, .917]). After audited corrections, 72.2% of all concordance lines were *Non-target* or *Unclassifiable*. This shows the limits posed by a lack of multi-modal context in text-only specialized corpora. Finally, we illustrate a proposed enriched transcription format for representing infant vocal activity, visible behavior, adult speech, caregiving action, and procedural context. We propose a privacy-preserving, source-decoupled derivative resource whose release would be subject to institutional review and platform/licensing constraints.

1 Introduction

The web has been used as a source for linguistic corpora for a number of years (Kilgarriff and Grefenstette, 2003; Baroni et al., 2006). More recently, web video has become a major source for audiovisual datasets: AudioSet uses human-labeled YouTube clips for sound-event recognition (Gemmeke et al., 2017). Elsewhere, YouTube-8M provides large-scale video labels and features (Abu-El-Haija et al., 2016), while VGGSound links audio classes to visible sound sources in YouTube clips (Chen et al., 2020). Another set, VoxCeleb, draws speaker data from online video (Nagrani et al., 2017), and still another, MOSI and MOSEI, use online opinion videos for multimodal sentiment and emotion analysis (Zadeh et al., 2016, 2018). Action-recognition datasets such as Kinetics and AVA show how web or already-circulating video can support fine-grained temporal annotation (Kay et al., 2017; Gu et al., 2018). These resources demonstrate the value of web audiovisual data, but they also foreground problems central to WaC research: platform bias, metadata sparsity, label noise, unstable availability, and the difficulty of documenting how source media become datasets. Infant-crying videos make these problems especially concrete because they combine highly variable user-generated recordings with sparse metadata, sensitive family and medical contexts, and identifiable minors.

A central difficulty in infant-cry research is that the cause of crying is often underspecified by the cry signal itself. Although cries reliably convey distress, specific eliciting causes are considerably harder to recover from acoustics alone. For example, Lockhart-Bouron et al. found that neither trained adult listeners nor machine-learning models could reliably distinguish hunger, discomfort, and isolation from cry acoustics; in their longitudinal corpus, cry causes were instead labeled primarily from the caregiver action that successfully stopped the crying, which agreed with the caregiver's prior attribution in 75% of cases (Lockhart-Bouron et al., 2023). This illustrates an important property of

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 331–346
October 29, 2026 ©2026 Association for Computational Linguistics

infant crying as interactional data: the eliciting context may be ambiguous at cry onset and may become clearer only through the surrounding situation and the caregiver's subsequent response. Feeding, repositioning, restoring proximity, removing a source of discomfort, or terminating an aversive stimulus can provide evidence that was not available from the cry alone.

At the same time, there is no broadly established data-driven typology spanning the range of everyday contexts in which young infants cry. Existing corpora use heterogeneous and often relatively narrow label sets, ranging from hunger and discomfort to clinically defined categories such as asphyxia or neurological conditions (Ji et al., 2021; Laguna et al., 2023). Reviews of infant-cry research also emphasize the scarcity of well-documented labeled data: many datasets have been collected in hospitals, NICUs, clinics, or controlled home studies, and many remain private or available only through restricted research access (Ji et al., 2021). This is particularly consequential in early infancy, where naturalistic, context-rich recordings are difficult to collect at scale. Our aim is therefore to develop an empirically grounded, context-sensitive, multimodal taxonomy of infant crying episodes from naturally occurring audiovisual material rather than to assume that discrete cry types can be read directly from acoustics alone. Web video is useful for this purpose because it can preserve the sequence linking an antecedent event or state, the infant's vocal and visible behavior, caregiver interpretation and action, and the subsequent continuation or resolution of distress. The resulting taxonomy is thus a typology of recoverable eliciting contexts, cues, and interactional evidence.

The contribution of this paper, in addition to an operational taxonomy of infant crying, is a compact data-centric description of how online audiovisual data can be used for empirical web-as-corpus research. We describe the analyzed sample, coding scheme, quality constraints, and validation procedure, together with a proposed privacy-preserving reuse model submitted for institutional ethical and data-protection review.

2 Corpus construction

Links to candidate videos were located between March and July 2026 on YouTube, Instagram, and TikTok. Undergraduate students contributed only to locating public links to videos of infants or babies crying and excluding links clearly concerning older children; they did not code or analyze the videos. No fixed search-term list or standardized retrieval procedure was imposed, and link-finding strategies could vary. The aim was broad, non-targeted discovery rather than targeted sampling of predefined cry contexts. Because retrieval was neither standardized nor randomized, the resulting sample should not be interpreted as a probability sample of infant-crying videos on these platforms.

The initial link set contained 325 candidate video records. Of these, 47 are marked as duplicates in the corpus table, and 28 additional non-duplicate records do not receive a retained cry-context label. Removing these 75 records leaves 250 unique classified videos. Video coding was carried out separately from link discovery. Initial context labels were assigned from the audiovisual record by the research team; Hexiang Wu, a graduate student in linguistics, also coded videos. The first author subsequently conducted an audit/validation pass in which the audiovisual context was re-examined and the original label could be confirmed, revised, or qualified with an explanatory note. This audit was not designed as a blinded, fixed-codebook inter-annotator agreement experiment: disagreements were also used to refine the coding guide. The 45 vaccination cases underwent a separate video-level context check and were all confirmed. Uploader-provided titles and descriptions were used for discovery but not as ground-truth labels. Corpus-level metadata used in the present analysis are deliberately sparse, containing context label and video duration; validation judgments are retained as a separate audit layer.

Compilations were retained when their crying episodes consistently represented the same cry context throughout; in such cases, the video received that single context label. Compilations for which a consistent cry context could not be established were not assigned a single-category label. The label *Multiple* was reserved for retained items in which more than one distinct cry context was independently supported. Table 1 summarizes the selection criteria.

The analyzed sample follows documentation principles developed for NLP and ML (machine learning) datasets (Bender and Friedman, 2018; Gebru et al., 2021; Paullada et al., 2021). For this domain, documentation must also encode what is deliberately *not* released and what metadata were *not* used. This is particularly important for web-

Criterion	Operational decision
Search protocol	Undergraduate students located public video links using self-selected search strategies; they did not code the videos. Older children were excluded.
Infant cry present	At least one audible infant cry or cry-like distress episode.
Context recoverable	Eliciting context visible/audible or explicitly attributed within the recording.
Audio usable	Cry pulses and adult speech audible enough for coding.
Video usable	Relevant visual context sufficiently observable for context coding.
Metadata used	Corpus-level analysis uses validated context label and video duration.
Sensitive content minimized	Source videos are not part of the proposed reusable corpus.

Table 1: Selection criteria for the web-derived infant crying corpus.

Field group	Examples
Corpus-level metadata	validated context label; video duration.
Discovery information	public links located through web search, not used as analytic labels.
Uploader metadata	titles/descriptions not used as ground-truth labels or analysis variables.
Proposed enriched transcript	infant cry notation; visible infant behavior; caregiving actions; speaker-attributed adult speech; salient scene/procedural events.
Proposed reuse	subject to ethical approval; de-identified, source-decoupled transcripts and annotations.

Table 2: Metadata and annotation fields used or proposed in the present corpus design.

derived audiovisual data, where large-scale collection practices have been criticized for weak consent, poor filtering, and opaque downstream reuse (Birhane and Prabhu, 2021). We therefore distinguish three layers: web-based discovery, audiovisual evidence used for annotation, and source-decoupled linguistic or multimodal representations proposed for reuse. In the current version, corpus-level metadata are restricted to context label and duration.

A further design decision is to keep the unit of discovery distinct from the unit of analysis. The web item is a video, but the linguistic or interactional evidence may be an infant vocalization, a caregiving action, an adult utterance, a visible infant behavior, or a procedural event such as an injection. This distinction makes it possible to use video-level labels for broad distributional work while representing selected interactions more richly in a source-decoupled transcript. The same design also avoids over-interpreting titles, uploader descriptions, or platform categories as linguistic metadata; these were discovery aids, not coding evidence. Platform of origin was not retained as an analytic variable in the sparse corpus-level table, so the present study does not report platform-stratified counts or treat platform as an explanatory factor.

3 Composition and taxonomy

The analyzed sample contains 250 unique classified videos. Current validation covers 230 (92.0%). Of these, 226 receive one of the current taxonomy labels and are summarized in Tables 3 and 4; four validation-covered records do not receive a current taxonomy label (three contain no crying and one remains unresolved between *Discomfort* and *Overstimulation*). Two validation judgments coded as proximity-seeking are treated as a descriptive subtype of *Fussiness*, consistent with the taxonomy below. Within the 226 labeled cases, 20 are *Unknown* and three are *Multiple*, leaving 203 substantive single-category audiovisual cases for the cross-modal comparison in Table 6. Labels describe the *eliciting context* supported by the recording, not an acoustic diagnosis. The category *Colic-like* is therefore used for audiovisual data, while *Colic* is retained below for web-text passages that explicitly use that term. The initial cry-context taxonomy was developed from existing research on infant crying, including work on acute pain, hunger, fussiness, discomfort, colic, and related behavioral states (Bellieni et al., 2004; Goberman and Whitfield, 2013; Parga et al., 2020; Laguna et al., 2023; Hammoud et al., 2024; Ji et al., 2021). It was subsequently refined through validation of the audiovisual labels: recurrent ambiguities and boundary cases motivated more explicit assignment criteria, including the use of *Colic-like* rather than *Colic* for audiovisual cases and stricter distinctions among *Gas*, *Discomfort*, *Fussiness*, *Birth*, and *Unknown*. Real-world cry detection likewise shows substantial in-the-wild noise and generalization challenges (Yao et al., 2020).

Existing infant-cry resources have largely been designed for acoustic classification. The Baby Chillanto database, for example, contains short cry segments grouped into clinically or behaviorally defined categories, while the crowdsourced Donate-a-Cry corpus contains parent-contributed audio recordings with contributor-supplied reason labels (Ji et al., 2021). The present resource differs in retaining the naturalistic audiovisual context surrounding crying episodes and in labeling the elicit-

Cry context	*n*	Mean length
Vaccination	45	1:43
Discomfort	44	2:16
Sleepiness	34	1:00
Fussiness	30	1:27
Hunger	13	1:27
Colic-like	12	2:05
Birth	8	3:46
Other Pain	8	2:08
Overstimulation	6	1:50
Gas	3	1:31
Multiple	3	10:45
Unknown	20	1:06
Total	226	1:50

Table 3: Validation-covered audiovisual cases assigned to the current cry-context taxonomy ($n = 226$).

Functional grouping	Corpus labels included	*n*
Homeostatic need	Hunger; Sleepiness	47
Immediate acute nociceptive	Vaccination; Other Pain	53
Externally elicited discomfort	Discomfort	44
Internal gastrointestinal / regulatory	Colic-like; Gas	15
Regulatory / interactional distress	Fussiness; Overstimulation	36
Perinatal transition	Birth	8
Unassigned / mixed context	Multiple; Unknown	23
Total		226

Table 4: Functional grouping of validated audiovisual cry-context labels.

ing context from multimodal evidence rather than treating isolated cry acoustics as the primary unit of analysis.

3.1 Validation-driven taxonomy refinement

Validation functioned not only as a reliability check but also as a mechanism for revising the coding scheme. The clearest case concerned *Sleepiness*. Among 40 videos initially labeled *Unknown* that received a validation judgment, 28 (70%) were reassigned to *Sleepiness*. This pattern exposed an omission in the first version of the taxonomy: visually recoverable sleep-state cues had not been operationalized sufficiently. The refined guide therefore treats a sleep transition, settling sequence, or partial-arousal state as relevant when supported by converging behavioral evidence such as repeated or prolonged eye closure, slow eyelid opening and closing, yawning, eye rubbing, or drowsy rearousal. Scene context such as a crib or bed can support the interpretation but is not sufficient by itself.

Validation also showed that *Fussiness* covers several recurrent forms of regulatory or protest distress rather than a single eliciting situation. We therefore use the label operationally, not as a claim that the crying has no real cause, is trivial, is under voluntary control, or ought to be ignored. Cry–fuss behavior can arise from interacting biological, environmental, and psychosocial factors, and the same overt crying may organize different caregiver responses (Douglas and Hill, 2013). Within the present material, recurrent subtypes include *goal-blockage protest*, for example when an object is removed or an ongoing activity ends; *proximity-seeking or separation-related protest*, when access to a caregiver is interrupted; and *socially elicited distress*, including contagious crying following another infant's cry. These remain descriptive subtypes rather than new top-level labels because the available evidence does not yet support stable independent categories.

This caution is especially important for sleep-related crying and so-called *cry-it-out* practices. Behavioral sleep interventions are heterogeneous, and evidence varies with age and method: trials of graduated-extinction approaches in older infants have not detected adverse effects on attachment or later emotional/behavioral outcomes, whereas work focused on the first six months has questioned the benefits of broadly applying such interventions and has emphasized the complexity of cry–fuss regulation (Gradisar et al., 2016; Price et al., 2012; Douglas and Hill, 2013). We therefore do not treat sleep training itself as evidence of neglect or trauma, nor do we equate any bedtime cry with *Sleepiness* or *Fussiness* in the audiovisual guide. Caregiver non-response, delayed response, or withdrawal may instead be represented as a separate interactional property when it is observable.

Qualitative inspection also surfaced a small set of contexts not well captured by the present top-level labels, including abandonment, prolonged unattended crying, frightening or rough caregiving, forced displacement, and distress following potentially traumatic events. Such cases motivate future distinctions concerning *caregiver responsiveness/adverse caregiving context*, but we avoid assigning clinical or legal labels such as *trauma* or *neglect* from a short web video alone. This distinction matters because adverse caregiving encompasses different forms of threat and deprivation that should not be collapsed into a single construct (Guyon-Harris et al., 2021); infant crying can also be a trigger for harmful caregiver responses,

making the caregiver response analytically distinct from the original cause of crying (Reijneveld et al., 2004). A further exploratory dimension concerns event continuity: many *Fussiness* cases involve ordinary but unwanted limitations on an ongoing goal or social-regulatory state, whereas *Overstimulation* more often follows abrupt, intense, novel, or expectancy-disrupting input. We treat these distinctions as hypotheses for future coding rather than additional current labels. Because the audit and taxonomy refinement were deliberately intertwined, the disagreement patterns remain evidence for codebook development rather than a fixed-codebook reliability test.

4 Web-text comparison: retrieval and cry-context recoverability

The audiovisual labels above are assigned from what can be seen and heard in each video. To test how far the same taxonomy can be recovered from linguistic material alone, we used the 52,268,286,493-word English enTenTen21 web corpus through Sketch Engine (Kilgarriff et al., 2014). We queried the corpus with the CQL expression `[lemma="baby"][lemma="cry"]`, which returned 14,046 concordance hits. No additional filters, sorting parameters, or randomization were applied. Sketch Engine therefore displayed the concordances in its default corpus order, and the present analysis uses the first 4,600 exported lines. The sample is consequently a contiguous, non-random subset of the query results. It is not a prevalence estimate, but a test of how often local web text supplies enough context to classify an infant-crying mention.

4.1 Contextual classification and validation

Each concordance was read from the exported local KWIC context (*Left* + *KWIC* + *Right*); no extended document context was consulted. Each line was judged individually with OpenAI GPT-5.6 Sol in ChatGPT. The procedure was deliberately contextual rather than keyword-rule based: the model determined whether an actual infant-crying event was described and, when possible, which eliciting context was supported by the passage. The output fields were `Classification`, `Confidence` (High/Medium/Low), and `Contextual_basis`, a short explanation of the evidence for the decision. The label set contained the ten shared cry contexts plus *Multiple*, *Unclassifiable*, and *Non-target*. *Multiple* was used only when more than one context was independently supported; *Unclassifiable* when an infant cry was present but its cause was not recoverable; and *Non-target* for figurative, quoted, imagined, fictional, device-related, lyric, or other non-event uses. Classification was performed interactively under a stable decision protocol rather than as a single one-shot API prompt; Appendix B records the operative rules.

Across the 4,600 lines, confidence was High for 4,094, Medium for 505, and Low for one. A human audit covered lines 1–460, exactly 10.0% of the classified set. The block was contiguous rather than randomly sampled, and the audit was corrective rather than blind: the auditor reviewed each KWIC line together with the assigned label, leaving `Validation` blank when the assignment was accepted and entering a correction or note otherwise. Of the 460 lines, 419 were accepted without an entry and 41 received a non-empty validation entry. For comparison with the declared label inventory, schema-equivalent wording was normalized (for example, *Unknown* to *Unclassifiable*); purely descriptive validator notes that did not constitute a taxonomy label were likewise collapsed to *Unclassifiable*. Two such entries were therefore equivalent to the original category, giving 421/460 normalized categorical agreements (91.5%; Wilson 95% CI [88.6, 93.7]). Cohen's κ was .885 (10,000-resample bootstrap 95% CI [.849, .917]). Because the auditor saw the assigned label, these are audit-agreement statistics rather than independent inter-annotator reliability estimates.

The audit covers the major retrieval decisions as well as substantive labels. Agreement was 151/164 (92.1%) for *Non-target*, 159/171 (93.0%) for *Unclassifiable*, 18/18 (100%) for *Multiple*, and 93/107 (86.9%) across the ten substantive labels (Table 5). Corrections were applied to the full counts: 1,931 *Non-target*, 1,388 *Unclassifiable*, 205 *Multiple*, and 1,076 substantive single-category cases. *Non-target* plus *Unclassifiable* account for 72.2% of the sample.

The audiovisual and web-text distributions reflect different modality-specific filters and are compared only within recoverable single-category cases. Vaccination and discomfort are more prominent audiovisually, where precipitating events can be directly visible; sleepiness, hunger, birth, other pain, and overstimulation are proportionally more frequent in web text, where writers often name an attributed cause. Fussiness is comparatively

Initial audit stratum	n	Agree	%
Non-target	164	151	92.1
Unclassifiable	171	159	93.0
Multiple	18	18	100.0
Substantive single category	107	93	86.9
All audited lines	460	421	91.5

Table 5: Human audit of the first 460 enTenTen21 concordance classifications. Agreement is reported after normalization to the declared web-text label inventory.

similar. Sleep-related discourse also exposes a boundary problem: *cry it out* or sleep-training references may describe a caregiving practice without establishing physiological sleepiness as the immediate cause. A post-hoc lexical check confirmed that such discourse spans several labels rather than forming a unique cry type. The comparison therefore concerns contextual recoverability, not population prevalence, and supports keeping *retrieval evidence*, *textual attribution*, and *audiovisual coding* distinct.

4.2 Causative constructions as a precision probe

A complementary query targeted the more explicitly causal construction *MAKE* + *baby/infant/newborn/neonate* + *CRY*. This query was used only as a precision probe and is not included in the 4,600-line distributional counts above. It returned 908 concordance lines. Although the construction should in principle make a causer easier to recover, 274 hits (30.2%) instantiated the recurrent *make Baby Jesus/Jebus cry* formula.[1] Related *Baby X* variants and other figurative or non-causal uses added further noise, and only 145 cases were provisionally mappable to a cry context. Under the contextual scheme used above, such formulaic cases are *Non-target* rather than infant-cry events.

The Baby Jesus/Jebus cases are analytically useful rather than merely accidental noise: they show that lexical proximity and even an apparently causal syntactic frame do not guarantee the intended referent or a literal crying event. Narrower WaC queries therefore still require contextual screening of formulaic, figurative, and pragmatic uses.

4.3 Refining *Other Pain*

Qualitative inspection of cases grouped under the broad *Other Pain* category during taxonomy development revealed recurrent subcontexts including teething or oral pain, circumcision and other medical procedures, burns or scalds, and general injury or trauma. These distinctions are not treated as separate top-level labels in the present taxonomy because the available examples remain sparse, but they indicate directions for future refinement of the category.

[1] *Jebus* is the intentional spelling occurring in the recurrent formula, not a typographical error.

5 Enriched representation for future interactional analysis

The video-level context labels distinguish eliciting situations that imply different candidate caregiver responses: hunger and sleepiness concern homeostatic regulation; vaccination and other pain provide acute nociceptive contexts; discomfort captures externally elicited non-acute physical discomfort; gas and colic-like crying capture internal gastrointestinal or regulatory contexts; and fussiness or overstimulation concern regulatory or interactional distress. This contextual layer can therefore support later analyses of how caregivers respond to a cry without treating the cry signal itself as a diagnosis. Vaccination work, for example, links caregiver emotional availability and soothing or distress-promoting behavior to infant pain regulation (Atkinson et al., 2015; Badovinac et al., 2018).

For selected future interactional analyses, we propose a source-decoupled representation aligning scene/procedural events, infant vocal activity, visible infant behavior, caregiving action, and minimized or paraphrased adult speech. Cry notation is descriptive rather than diagnostic: `@_C/@_c` mark relatively louder/softer pulses, `_phon/_dys` phonated versus dysphonic quality, and `_ex/_insp` expiratory versus inspiratory phase. Table 7 gives a synthetic example solely to demonstrate the representation; it does not reproduce any individual source interaction.

6 Quality control and audit trail

After duplicate-marked and non-retained rows are excluded, 250 unique classified videos remain. Current validation covers 230/250 (92.0%): 200 retained records have a non-empty explicit audit/validation judgment, including 15 vaccination cases, and the remaining 30 vaccination videos underwent the separate video-level context check described above. For agreement calculation, trivial spelling/spacing differences and parenthetical explanatory notes were normalized. Before adjudication and taxonomy refinement, the initial and

Cry context	Audiovisual		Web text		Δ pp
	n	%	*n*	%	
Vaccination	45	22.2	10	0.9	-21.2
Discomfort	44	21.7	120	11.2	-10.5
Fussiness	30	14.8	176	16.4	+1.6
Colic-like / Colic	12	5.9	87	8.1	+2.2
Hunger	13	6.4	137	12.7	+6.3
Birth	8	3.9	110	10.2	+6.3
Other Pain	8	3.9	93	8.6	+4.7
Overstimulation	6	3.0	85	7.9	+4.9
Gas	3	1.5	1	0.1	-1.4
Sleepiness	34	16.7	257	23.9	+7.1
Total	203	100.0	1,076	100.0	

Table 6: Distribution of the ten shared contexts in validated audiovisual cases and single-category web-text classifications. Audiovisual *Colic-like* is compared with web-text *Colic*. Percentages are within classified cases, not population prevalence estimates.

Tier	Synthetic representation
PROC	((Injection occurs.))
BABY	`@_C_phon_ex` → `@_C_dys_ex`
VIS	((Infant withdraws and raises arms.))
ADULT	((Caregiver moves closer and provides soothing touch.))

Table 7: Synthetic illustration of the proposed source-decoupled representation. No row reproduces an identifiable source interaction.

validation labels agree in 156/230 cases (67.8%; Cohen's $\kappa = .636$). Restricting the denominator to validation-covered records whose initial label was substantive gives 145/188 agreement (77.1%; $\kappa = .732$).

These values are best interpreted as *audit statistics*, not as reliability estimates for a frozen final codebook. The validation pass was used both to check individual labels and to sharpen category boundaries, and it was not designed as a blinded fresh annotation exercise. Final distributional tables use the validated/adjudicated labels. Of the 230 validation-covered records, 226 receive a current taxonomy label; three no-cry cases and one unresolved *Discomfort*/*Overstimulation* case do not enter the taxonomy distributions. The two proximity-seeking judgments are nested under *Fussiness*. Exclusion of *Unknown* and *Multiple* therefore leaves 203 single-category audiovisual cases compared with web text. A conventional category-level inter-annotator reliability analysis would require a fresh pass in which multiple coders independently apply the finalized guide.

The disagreement pattern was itself analytically useful: it identified categories whose operational criteria required revision rather than treating all disagreements as interchangeable coding error. The systematic reassignment of initially *Unknown* cases to *Sleepiness*, and the recurrent subtypes observed within *Fussiness*, are described in Section 3.1.

7 Ethics, data protection, and proposed reuse

The source material consists of publicly accessible videos on YouTube, Instagram, and TikTok. The videos were viewed on their source platforms for research coding; they were not downloaded or retained as audiovisual files in the present phase. Because the recordings involve infants, caregivers, and, in some cases, medical procedures, birth, or visible distress, public accessibility is not treated as sufficient justification for unrestricted reuse or redistribution. The paper therefore reports aggregate results and uses synthetic rather than case-level audiovisual illustrations.

The terms of the three source platforms were checked on 20 September 2026. YouTube restricts downloading, reproduction, distribution, modification, and automated access except where permitted by the service, rights holders, or applicable law.[2] Instagram prohibits unauthorized access or collection and requires respect for third-party intellectual-property rights.[3] TikTok's EEA terms restrict unapproved automated extraction and make some reuse, including downloading, dependent on platform features and settings.[4] Public availability is therefore not treated as a blanket licence for local acquisition or redistribution.

The present study therefore keeps the obser-

[2]YouTube Terms of Service, "Permissions and Restrictions": https://www.youtube.com/t/terms.

[3]Instagram Terms of Use: https://help.instagram.com/581066165581870.

[4]TikTok Terms of Service (EEA/UK/Switzerland): https://www.tiktok.com/legal/page/eea/terms-of-service/en.

vational analysis distinct from a possible future preservation phase. Subject to favorable institutional ethical and data-protection review, and only where the applicable platform terms, rights-holder permissions/licences, or legal exceptions permit it, a future phase could temporarily acquire authorized source material solely to create privacy-preserving derivatives. Proposed safeguards include neutral or synthetic facial rendering driven by landmarks/action units, removal or inpainting of identifying backgrounds and visible text, voice de-identification or replacement by transcription, metadata stripping, and re-identification testing. A derivative would not be described as anonymous unless that standard could be demonstrated; otherwise audiovisual derivatives would not be openly released and access would remain restricted.

8 Discussion

The validation results show why taxonomy development and reliability assessment are not separate stages in this type of corpus. The large movement from *Unknown* to *Sleepiness* did not simply add another label; it revealed that the first coding guide had under-specified a visually available evidential domain. Once sleep-state transitions and converging cues such as repeated eye closure or partial arousal were made explicit, a substantial set of previously residual cases became classifiable. Conversely, the refinement of *Fussiness* showed that a useful top-level category can remain internally heterogeneous. Goal-blockage protest, proximity-seeking or separation-related protest, and socially elicited distress can be kept together at the present level while remaining available as analytically meaningful subtypes.

The same validation process suggests a potentially useful cross-cutting distinction between ordinary regulatory frustration and abrupt event discontinuity. Many fussiness cases concern a familiar situation becoming temporarily undesirable or unavailable; overstimulation more often involves sudden, intense, novel, or expectancy-disrupting input. Other pain can share this abruptness while remaining distinguished by an identifiable nociceptive event. Future work can test whether this event-continuity dimension improves annotation reliability or predicts differences in infant or caregiver response before it is incorporated into the coding scheme.

Uploader metadata may also be analytically useful as a separate attribution layer. One audiovisual case remained *Unknown* from the recording itself but was described in uploader metadata as hunger-related. More generally, terms such as *fussy*, *self-soothing*, and *cry it out* can encode a caregiver's interpretation of the infant and a stance toward appropriate response rather than a directly observable cause. This is particularly consequential online, where sleep-training and responsiveness practices are publicly debated and the same crying episode may be framed as sleepiness, protest, a need for proximity, or behavior that should be allowed to resolve without intervention. Rather than using uploader wording to overwrite audiovisual evidence, we therefore treat metadata as an attribution layer that can itself be studied.

9 Limitations

The audiovisual sample is exploratory rather than representative. Candidate links were located through unconstrained, non-random discovery on YouTube, Instagram, and TikTok, so inclusion reflects public posting and platform search/ranking practices. Platform provenance was not retained analytically, preventing platform-stratified counts. No French-language videos occur in the sample, but this does not imply international representativeness. Demographic and clinical information is usually unavailable, recording conditions vary, and source availability changes over time. The distributions therefore describe this web-visible sample rather than population prevalence.

The audiovisual validation procedure also has an important design limitation: it served both as a label audit and as a mechanism for refining the taxonomy, rather than as blinded application of a frozen codebook. The reported values are therefore pre-adjudication audit statistics, not final inter-annotator reliability. Several categories remain small, especially *Gas*, *Birth*, *Other Pain*, and *Overstimulation*. A stronger test would require a fresh independent coding pass using Appendix A, with category-level agreement and a confusion matrix. *Fussiness* is likewise an operational, heterogeneous context label, not a claim that distress is trivial or lacks a cause.

The web-text analysis has different constraints. The 4,600 lines are the first contiguous query block rather than a random sample, and classification uses exported KWIC context rather than full documents. The 10% audit is corrective, not blind,

so 91.5% agreement and $\kappa = .885$ quantify audit agreement rather than independent reliability. The *cry-it-out*/Ferber cases further show that parenting terminology may describe caregiving practice or stance without uniquely identifying the infant's immediate state, particularly for *Sleepiness* and *Fussiness*.

Exact audiovisual reconstruction is intentionally limited by the privacy-preserving design. Source media and identifiable case-level material are not released, and uploader metadata were not systematically retained as an analytic layer. Reproducibility is therefore strongest for the coding scheme, aggregate counts, web-text query, and audit protocol. Future work should retain ethically permissible provenance variables and independently evaluate the frozen taxonomy.

10 Conclusion

We presented an analyzed sample of 250 unique web-derived infant crying videos from YouTube, Instagram, and TikTok, an operational context taxonomy, and a validation audit covering 230 videos (92.0%). Of these, 226 receive current taxonomy labels, and 203 substantive single-category cases enter the cross-modal comparison. The audit provides pre-adjudication agreement statistics while also documenting how the taxonomy was refined: in particular, it exposed the under-specification of sleep-related visual evidence and revealed recurrent regulatory/protest subtypes within *Fussiness*. Reported audiovisual distributions use validated labels only.

A supporting enTenTen21 analysis shows that textual cry context can be recovered at scale through contextual judgment, but also that many web concordances are non-target or leave the eliciting cause unspecified. A human audit of 460 lines (10.0%) covered the full output inventory rather than only substantive cases and yielded 91.5% normalized categorical agreement ($\kappa = .885$). The cross-modal comparison further shows why textual attribution and audiovisual evidence should remain distinct: a cause may be linguistically named in text yet visually unavailable, while an event such as an injection may be directly observable without being verbally identified.

Finally, the synthetic vaccination sequence demonstrates the proposed enriched transcription format without reproducing an individual source interaction. The format keeps contextual labels distinct from observable infant vocal activity, visible behavior, adult speech, and caregiving action. For Web-as-Corpus research, the study therefore illustrates a workflow in which broad discovery, audiovisual coding, validation-driven taxonomy refinement, contextual web-text classification, multimodal transcription, and release constraints are documented separately.

Acknowledgments

We thank Alex Aguiar, Telmen Altanbagana, Darline Aurivel, Anamika Das, Cassidy Desousa, Alexandra Gelabert, Juliette Guiet Dupre, Apolline Huet, Lila Parent, Makizhiny Paul, Danica Penaflorida, Charlene Pires, Lison Rose, Clara Silvestre, Niruya Sivanand, Ines Torchon, Paris-Cité undergraduate students for their assistance with locating publicly accessible video links during the initial corpus-discovery stage and for interesting discussions. Their contribution to corpus compilation was limited to identifying candidate links; video coding and analysis were carried out separately by the research team.

This work was supported by the ANR CODIM Project (https://www.codim-project.org/) and by the investment program "France 2030" launched by the French Government and implemented by the University Paris Cité as part of its program "Initiative d'excellence" IdEx with the reference (ANR-10-IDEX-0001).

AI use statement

OpenAI ChatGPT was used during manuscript preparation for drafting, restructuring, and language editing. The authors reviewed and revised the resulting text and are responsible for the final content. The separate use of GPT-5.6 Sol for concordance classification is described in Section 4 as part of the study methodology.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. In *arXiv preprint arXiv:1609.08675*.

Nicole H. Atkinson, Hannah Gennis, Nicole M. Racine, and Rebecca R. Pillai Riddell. 2015. Caregiver emotional availability, caregiver soothing behaviors, and infant pain during immunization. *Journal of Pediatric Psychology*, 40(10):1105–1114.

Shaylea Badovinac, Hannah Gennis, Rebecca Pillai Riddell, Hartley Garfield, and Saul Greenberg. 2018. Understanding the relative contributions of sensitive and insensitive parent behaviors on infant vaccination pain. *Children*, 5(6):80.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2006. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Carlo V. Bellieni, Renata Sisto, Duccio M. Cordelli, and Giuseppe Buonocore. 2004. Cry features reflect pain intensity in term newborns: An alarm threshold. *Pediatric Research*, 55(1):142–146.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1547.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 721–725.

Pamela S. Douglas and Peter S. Hill. 2013. Behavioral sleep interventions in the first six months of life do not improve outcomes for mothers or infants: A systematic review. *Journal of Developmental & Behavioral Pediatrics*, 34(7):497–507.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780.

Alexander M. Goberman and Jason A. Whitfield. 2013. Acoustics of infant pain cries: Fundamental frequency as a measure of arousal. *Perspectives on Speech Science and Orofacial Disorders*, 23(1):18–26.

Michael Gradisar, Kate Jackson, Nicola J. Spurrier, Joyce Gibson, Justine Whitham, Anne Sved Williams, Robyn Dolby, and David J. Kennaway. 2016. Behavioral interventions for infant sleep problems: A randomized controlled trial. *Pediatrics*, 137(6):e20151486.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056.

Katherine L. Guyon-Harris, Kathryn L. Humphreys, and Charles H. Zeanah. 2021. Adverse caregiving in early life: The trauma and deprivation distinction in young children. *Infant Mental Health Journal*, 42(1):87–95.

Mohammed Hammoud, Melaku N. Getahun, Anna Baldycheva, and Andrey Somov. 2024. Machine learning-based infant crying interpretation. *Frontiers in Artificial Intelligence*, 7:1337356.

Chunyan Ji, Thosini Bamunu Mudiyanselage, Yutong Gao, and Yi Pan. 2021. A review of infant cry analysis and classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):8.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: Ten years on. *Lexicography*, 1:7–36.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. In *Computational Linguistics*, volume 29, pages 333–347.

Ana Laguna, Sandra Pusil, Àngel Bazán, Jonathan Adrián Zegarra-Valdivia, Anna Lucia Paltrinieri, Paolo Piras, Clàudia Palomares i Perera, Alexandra Pardos Véglia, Oscar Garcia-Algar, and Silvia Orlandi. 2023. Multi-modal analysis of infant cry types characterization: Acoustics, body language and brain signals. *Computers in Biology and Medicine*, 167:107626.

Marguerite Lockhart-Bouron, Andrey Anikin, Katarzyna Pisanski, Siloé Corvin, Clément Cornec, Léo Papet, Florence Levréro, Camille Fauchon, Hugues Patural, David Reby, and Nicolas Mathevon. 2023. Infant cries convey both stable and dynamic information about age and identity. *Communications Psychology*, 1(1):26.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech*, pages 2616–2620.

Joanna J. Parga, Sharon Lewin, Juanita Lewis, Diana Montoya-Williams, Abeer Alwan, Brianna Shaul, Carol Han, Susan Y. Bookheimer, Sherry Eyer, Mirella Dapretto, Lonnie Zeltzer, Lauren Dunlap, Usha Nookala, Daniel Sun, Bianca H. Dang, and Ariana E. Anderson. 2020. Defining and distinguishing infant behavioral states using acoustic cry analysis: Is colic painful? *Pediatric Research*, 87(3):576–580.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Anna M. H. Price, Melissa Wake, Obioha C. Ukoumunne, and Harriet Hiscock. 2012. Five-year follow-up of harms and benefits of behavioral infant sleep intervention: Randomized trial. *Pediatrics*, 130(4):643–651.

Sijmen A. Reijneveld, Marcel F. van der Wal, Emily Brugman, Remy A. Hira Sing, and S. Pauline Verloove-Vanhorick. 2004. Infant crying and abuse. *The Lancet*, 364(9442):1340–1342.

Xuewen Yao, Megan Micheletti, Mckensey Johnson, Edison Thomaz, and Kaya de Barbaro. 2020. Infant crying detection in real-world environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3).

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. In *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246.

A Cry-context coding guide

The labels classify the *eliciting context of crying*, not an acoustic cry type or clinical diagnosis. Evidence is ranked as follows: (1) participant attribution or in-video text; (2) a directly observable proximal event or condition; (3) converging infant behavior; and (4) cry characteristics as supporting evidence only. Uploader titles, descriptions, tags, and search terms are discovery metadata and do not determine the audiovisual label. The scheme is informed by work distinguishing acute pain, hunger, fussiness, discomfort, colic, and related contexts (Bellieni et al., 2004; Goberman and Whitfield, 2013; Parga et al., 2020; Laguna et al., 2023; Hammoud et al., 2024; Ji et al., 2021). The present taxonomy additionally separates GAS from externally elicited DISCOMFORT, since gastrointestinal gas or straining need not be painful and is not necessarily caused by an external event.

B Web-text contextual classification protocol

For each enTenTen21 line, only the exported *Left* + *KWIC* + *Right* context was considered. The classifier was instructed to make a contextual judgment rather than match keywords: first determine whether the passage describes an actual infant-crying event; assign NON-TARGET to figurative, fictional, quoted, imagined, device-related, lyric, or otherwise non-event uses; assign UNCLASSIFIABLE when an actual cry is present but its eliciting context cannot be recovered; use MULTIPLE only when more than one context is independently supported; otherwise select one of the ten shared cry-context labels. Causes were not inferred from setting or co-occurrence alone (e.g., feeding did not by itself license HUNGER, and a distressing scene did not by itself license OTHER PAIN). Each decision returned a label, High/Medium/Low confidence, and a short contextual justification.

Label	Assign when	Do not assign from
Vaccination	A visible injection/vaccination is temporally linked to crying, or participant speech/in-video text explicitly identifies it.	Clinic setting or a needle elsewhere in the scene alone.
Other Pain	A painful procedure, injury, or clearly painful pressure/compression is visibly or explicitly established and linked to the cry (e.g. heel prick, blood draw, injury).	Intensity, roughness, ordinary handling, or presumed pain without an identifiable painful event.
Birth	The cry is a spontaneous first/early neonatal cry in the immediate post-delivery transition. Routine cleaning may coexist when it is not the elicitor.	If active chest pounding, hard rocking, or another intervention is used to elicit a delayed cry, code the immediate elicitor as DISCOMFORT or OVERSTIMULATION. An unresponsive infant with no cry is not a crying case.
Hunger	Participant speech/in-video text attributes hunger, or a clear pre-feeding sequence plus converging rooting/searching/sucking behavior supports it; feeding that promptly resolves the episode may corroborate.	A bottle/breast/feed somewhere in the video, post-feeding crying, or uploader metadata alone.
Sleepiness	Tiredness/sleep is explicitly attributed or a clear sleep transition, partial-arousal state, or settling sequence is supported by drowsiness, yawning, eye-rubbing, repeated eye closure, etc.	Night-time, crib/bedtime, eventual sleep, or a sleep-training/cry-it-out frame alone.
Gas	Gas, trapped air, difficulty burping, reflux, constipation/straining, or a related gastrointestinal process is explicitly attributed or strongly established, with converging behavior such as leg flexion, straining, regurgitation, or attempts to burp or pass gas. Pain is not required.	Burping, feeding, passing gas, or tummy touching alone; do not equate gas with colic-like crying or with externally elicited discomfort.
Discomfort	An externally caused, non-acute physical condition or handling is clearly implicated: wet/soiled diaper, bathing/cleaning, uncomfortable position or clothing, temperature exposure, skin irritation, restrictive mittens, restraint, or a sustained difficult posture.	Internal gastrointestinal states such as gas, reflux, or straining; or diaper change, bathing, clothing, mittens, handling, or posture merely being present without evidence that the external condition contributes to crying.
Overstimulation	A sudden/intense sensory or social stimulus immediately precedes the cry; startle, withdrawal, turning away, escalation with continued stimulation, or relief when it stops may support the label.	Background television/music, ordinary conversation, people in the room, or ambient noise without a temporal relation.
Fussiness	Regulatory/protest distress after specific homeostatic, gastrointestinal, externally elicited discomfort, and nociceptive causes are unsupported; subtypes may include goal blockage, proximity-seeking, and socially elicited distress.	Not an unexplained-cry default. The label does not imply triviality or endorse non-response; do not assign from sleep-training framing or caregiver non-response alone.
Colic-like	Prefer explicit participant/in-video attribution to colic. Without it, require sustained or recurrent difficult-to-console crying in a sufficiently extended recording, with other specific contexts excluded; otherwise cue-free clips under 30 s are UNKNOWN.	Cry acoustics, intensity, apparent gas, or unexplained crying alone. The label is descriptive, not a clinical diagnosis.
Multiple	Two or more different contexts independently satisfy their own criteria within the same item.	Uncertainty between alternatives; ambiguous alternatives are UNKNOWN, not MULTIPLE.
Unknown	No category meets its evidential criterion, or cues conflict so that the eliciting context cannot be assigned confidently.	A "most likely" guess made only to avoid an unknown label.

Table 8: Concise operational guide for audiovisual cry-context coding.

C Sample of the web-text classification file

Table 9 reproduces the first 100 rows of the 4,600-line enTenTen21 classification file with all exported fields. The *Left* and *Right* columns are reproduced in full rather than abridged. A blank *Validation* cell means that the human audit accepted the assigned classification unchanged.

Table 9: First 100 rows of the 4,600-line enTenTen21 contextual-classification file.

Line	Reference	Left	KWIC	Right	Classification	Conf.	Contextual basis	Validation
1	chs.ca	. </s><s> However, additional transmitters can be added to receive alerts for the telephone/TTY ringing (BE1023),	baby crying	(BE1290) and someone at the door (BE1123 or BE1240). </s><s> Silent Call Medallion Series </s><s> What it is: Smoke detector with	Non-target	High	Device/accessibility example; no infant crying event is being described.	
2	hnn.us	you are nothing but a victim and grow up, get educated and be pro-active. </s><s> This country is like living with a bunch of adult	babies crying	for Mommy Government to take care of them and blaming all of societies ills on Capitalism. </s><s> Why don"t these whiners get on	Non-target	High	Metaphorical use: adults compared to crying babies.	
3	hnn.us	this brutal system. </s><s> A mother with a young baby falls in the stifling heat and both are beaten mercilessly by a guard. </s><s> The	baby cries	, but the mother knows to stay silent. </s><s> In this camp, fewer than half the prisoners will survive. </s><s> Even the commandant	Other Pain	High	The baby and mother have just been beaten by a guard.	
4	cmc.ie	through repetition, and these are all sounds thematically relevant to the Wake. </s><s> The sounds of water, church bells,	babies crying	and children laughing, sea gulls, thunder, snippets of voices although half-overheard, work their way through the	Non-target	High	Generic list of sounds in an artistic work; no infant event/cause.	
5	nhs.uk	Ajitomi, Daihei Shiohama. </s><s> W3C"s hosting of this group does not imply endorsement of the activities. </s><s> Colic is when a	baby cries	a lot but there"s no obvious cause. </s><s> It"s a common problem that should get better on its own. </s><s> But call NHS 111 or see a GP if	Colic	High	Explicitly defines colic as prolonged crying without obvious cause.	
6	nhs.uk	problem that should get better on its own. </s><s> But call NHS 111 or see a GP if you"re worried. </s><s> Check if your baby has colic </s><s> All	babies cry	, but your baby may have colic if they cry more than 3 hours a day, 3 days a week for at least 1 week. </s><s> Immediate action	Colic	High	Context is a colic diagnostic description.	
7	nhs.uk	week for at least 1 week. </s><s> Immediate action required: Go to A&E or call 999 if: </s><s> your baby has a weak or high-pitched cry </s><s> your	baby"s cry	does not sound like their normal cry </s><s> You know better than anyone else what your child is usually like. </s><s> Trust your	Unclassifiable	High	Abnormal cry is flagged medically, but no cause is given.	Medically concerning cry (abnormal cry)
8	hse.ie	old. </s><s> A baby has colic if they cry: </s><s> for more than 3 hours a day </s><s> Colic can be very upsetting for parents. </s><s> It is hard to see your	baby crying	so intensely, and it can seem like your baby is in pain. </s><s> Signs and symptoms of colic </s><s> Signs and symptoms that your baby has	Colic	High	Explicit signs/symptoms of colic.	
9	u2.com	the world but you knew better And that all we have is immortality Don"t close your eyes Love and love is all we have left A	baby cries	on a doorstep Love is all we have left Love and love is all we have left You argue because you can"t accept Love is all we have	Non-target	High	Song lyric rather than a described infant-crying episode.	
10	u2.com	stars in her eyes So many stars So many ways of seeing Hey this is no time not to be alive Love and love is all we have left A	baby cries	on a doorstep Love is all we have left Love and love is all we have left The only thing that can be kept Love is all we have left </s>	Non-target	High	Song lyric rather than a described infant-crying episode.	
11	hse.ie	that you experience can include disbelief, anger, guilt and grief. </s><s> It is quite common to think you can hear your	baby crying	. </s><s> Your healthcare team is there to support you through this bereavement. </s><s> They will give information to help you to make	Non-target	High	Bereaved person thinks they hear the baby; no actual crying event.	
12	om.org	her faith in Jesus as her Saviour. </s><s> Upon hearing the prayer, Rada announced, "I want to pray this prayer!" </s><s> Then her	baby cried	; the phone rang; everything happened at once. </s><s> But once it calmed down, Rada again said that she wanted to pray the prayer	Unclassifiable	High	An actual baby cries, but the cause is not stated.	
13	tnq.ca	. </s><s> He sits in the narrow hallway, warming his nose between his paws, and looks up at Pema with big, mournful eyes. </s><s> The	baby"s cry	could break your heart. </s><s> Her head is flung back in abandon, her tiny lungs unequal to her despair. </s><s> It"s not colic or hunger	Unclassifiable	Medium	Cry is intense and hunger/colic are ruled out, but no alternative cause is supplied.	
14	rt.com	, something often called "treasonous" ... They – wait for it – think for themselves! </s><s> (Glass shatters. </s><s> Woman screams. </s><s>	Baby cries	.) </s><s> Nothing is more frightening for our endless war machine than a military grunt who thinks for him or herself. </s><s> They"re	Non-target	High	Parenthetical sound effect used rhetorically.	
15	cbc.ca	<s> Protesters were caught between the Mexican and U.S. authorities. </s><s> A young woman fell to the ground unconscious, and two	babies cried	, tears streaming from the gas, a Reuters witness said. </s><s> "They want us to wait in Mexico but I for one am desperate. </s><s> My	Other Pain	High	Babies are crying while exposed to tear gas.	
16	cbc.ca	by what has stolen from her. </s><s> Sometimes she hears her baby in her sleep. </s><s> "I"ve had a few dreams ... where you could hear a	baby crying	or you could have a sense of a baby," she says. </s><s> "The first time I had it I didn"t know if it was a boy or a girl. </s><s> And then another	Non-target	High	Dream/imagined baby crying.	
17	kk.org	Nasal Aspirator ($19) "Our kids are too big for these now, but when they were babies, these things were super helpful. </s><s>	Babies cry	for four reasons: hunger, tired, diaper, gas. </s><s> The Windi helps with the fourth. </s><s> When a kid is sick and can"t breath well	Multiple	High	Explicitly lists hunger, tiredness, diaper and gas as possible causes.	
18	imt.ie	from the Priory Hall residents. </s><s> The IMO should be pushing out NCHDs who are young mothers and showing pictures of their	babies crying	and telling stories over and over on Joe Duffy and the like about young NCHDs with pretty faces forced to work the day	Non-target	Medium	Rhetorical/media example rather than a specific infant-crying context.	
19	imt.ie	and nourish your offspring... and this is all the consolation you need for that 4am feed, yet again. </s><s> But what if your	baby"s cries	weren"t as innocent as you think? </s><s> New research, published in the journal Evolution, Medicine, and Public Health,	Unclassifiable	Medium	Discussion of infant crying research, but no cause for the cry is given.	
20	dw.com	don"t breathe air in the womb) and the cycle starts all over again. </s><s> It"s a funny thought. </s><s> And a bit gross. </s><s> 10. </s><s> If you hear a	baby cry	, your breastmilk may appear at the ready ... </s><s> ... even if it"s not your own child crying. </s><s> Midwife Peter Wolf describes	Unclassifiable	High	Real baby crying is invoked, but only as a trigger for milk let-down; cause unknown.	
21	dw.com	woman with her skull cracked open. </s><s> Haifee was about to leave her body in the debris to look for survivors, until he heard a	baby crying	from underneath her. </s><s> When he turned over the corpse, he found two-month old Kareem Abdullah covered in dust and	Unclassifiable	Medium	Infant is found under a corpse in disaster debris; distress is clear, but the immediate cause of crying is not established.	
22	imt.ie	. </s><s> Pre-natal education on newborn crying behaviour was important, she said. </s><s> Maternal depression was not the cause of	babies crying	, said Dr Sheridan-Pereira, but parents could end up depressed. h4. </s><s> Multiple aetiology Ms Roberta McCarthy, neonatal	Multiple	Medium	The passage explicitly frames newborn crying as having multiple aetiologies.	
23	bgr.in	instance, Galaxy devices can distinguish sounds more accurately, thanks to an improved AI. </s><s> So, if the device detects a	baby crying	or a dog barking in everyday surroundings, it will send an alert to the user"s smartphone. </s><s> Users can listen to the	Non-target	High	AI sound-detection example.	
24	nj.gov	and Science in TIMSS 2015 </s><s> Part of </s><s> Share </s><s> These materials can be downloaded or ordered through DCF Publications. </s><s> When a	Baby Cries	(Shaken Baby Syndrome) </s><s> When a Baby Cries (Shaken Baby Syndrome) - Campaign warns parents/caregivers to never shake a	Unclassifiable	Medium	Campaign concerns coping with a crying baby, but no cause of crying is specified.	
25	nj.gov	<s> These materials can be downloaded or ordered through DCF Publications. </s><s> When a Baby Cries (Shaken Baby Syndrome) </s><s> When a	Baby Cries	(Shaken Baby Syndrome) - Campaign warns parents/caregivers to never shake a baby because it can cause serious brain	Unclassifiable	Medium	Campaign concerns coping with a crying baby, but no cause of crying is specified.	
26	h4.org	to Big Orange Balls for yelling at him for taking circle beer when he was only trying to do his reboot down down and made his	baby cry	. </s><s> On On Pipes Space City RA (Houston) Hamersley Hash Roving Ambassador </s><s> Cashier"s Office & Student Account Enquiries </s>	Non-target	Low	No clear infant context; likely figurative/opaque use in a hash-related passage.	

Continued on next page

Table 9 – continued from previous page

Line	Reference	Left	KWIC	Right	Classification	Conf.	Contextual basis	Validation
27	nj.gov	and available by completing the PUBLICATIONS ORDER FORM. </s><s> Some materials are also available for downloading. </s><s> When a	Baby Cries	(Preventing Shaken Baby Syndrome) This brochure gives you tips on coping with a crying baby and warns of the dangers of	Unclassifiable	Medium	Brochure is about coping with a crying baby; cause is not specified.	
28	rc.org	also ask questions, such as, "Who in your daily life can listen to you effectively?" </s><s> "What happens for you when you hear a	baby crying	?" </s><s> "Have you noticed that people use your attention to get their thoughts straight?" </s><s> Then I learn from the answers. </s><s> The	Non-target	High	Generic interview question about hearing a baby cry.	
29	cbc.ca	instructions and how to care for the baby and mom before the ambulance reaches them," said St. Onge. </s><s> As soon as you hear	baby crying	or screaming, that"s a good sign.- Lyndsie St. Onge </s><s> St. Onge said her priority was to make sure the baby was breathing and	Birth	High	Crying is used as a sign that a newly delivered baby is breathing.	
30	imt.ie	, our parents were told to put them to bed and let them bawl. </s><s> It was tough love. </s><s> Later on, we were informed that leaving	babies cry	was tantamount to assault. </s><s> {openx:269} The last generation believed in stoically in making the best of a bad lot. </s><s> The	Sleepiness	Medium	Crying occurs in the context of putting babies to bed/cry-it-out sleep practice.	
31	nm.org	and so overtired," Ideran says. </s><s> "After a few sessions, parents often report that they have a different baby. </s><s> The	baby cries	less, and eats and sleeps at more regular intervals." </s><s> Ideran teaches the babies different ways to self-calm. </s><s> "Some	Sleepiness	High	Baby is explicitly described as overtired.	
32	vch.ca	and play with your baby. </s><s> Always keep one hand on the baby and never turn your back while they are on the changing table. </s><s> All	babies cry	. </s><s> It"s the way they communicate. </s><s> But sometimes you"ve tended to all their needs and your baby still won"t stop crying. </s><s> If	Fussiness	High	Needs have been tended to but the baby still will not stop crying.	Unknown
33	imt.ie	through an ultrasound scan, sounds of blood supply, of women whispering or shouting during delivery and newborn	babies crying	. </s><s> Dr Cohen is a French and a British citizen. </s><s> He first studied piano, classical music and jazz and then spent time at the "	Birth	High	Newborn babies crying during delivery.	
34	ie.edu	behavior on video calls. </s><s> For example, managers might view home background noise, such as a dog barking or	baby crying	, to imply that others are not fully present. </s><s> A misplaced webcam can result in a worker being interpreted as evasive or	Non-target	High	Background-noise example in discussion of video calls.	
35	nm.org	information will help you develop realistic expectations and feel confident about caring for your new baby. </s><s> Crying </s><s>	Babies cry	as a way to communicate hunger, soiled diaper, overstimulation, pain, frustration and even loneliness, and in time	Multiple	High	Explicitly lists hunger, soiled diaper, overstimulation, pain, frustration and loneliness.	
36	imt.ie	discount launched in India: Price, specifications, features </s><s> The neverending debate just got more fuel: to let	babies cry	, or not to </s><s> Parenting expert Penelope Leach says it is now a fact that letting babies cry is potentially harmful to their	Unclassifiable	Medium	General cry-it-out debate; the immediate cause of crying is not stated.	
37	imt.ie	debate just got more fuel: to let babies cry, or not to </s><s> Parenting expert Penelope Leach says it is now a fact that letting	babies cry	is potentially harmful to their developing brains. </s><s> Gina Ford is expected to pounce with more facts soon. </s><s> A Guardian	Unclassifiable	Medium	General cry-it-out debate; the immediate cause of crying is not stated.	Fussiness
38	hse.ie	as well as the impact of stress on pain sensitivity and spinal circuitry. </s><s> Communication skills change as we grow –	babies cry	, while older children can be more specific in their communication. </s><s> There are many ways to communicate including the	Unclassifiable	High	Generic statement that babies cry to communicate; no specific cause.	
39	w3.org	: Captioning Tools. </s><s> What to Transcribe </s><s> Generally, you transcribe all speech and relevant non-speech sound (such as:	baby cries	, fireworks going off, horse hoofs approaching). </s><s> Keep in mind that the main purpose is to provide the information that	Non-target	High	Transcription guideline example.	
40	nj.gov	, the company noted that both the series will </s><s> It"s normal for babies to cry, even when you are trying to comfort them. </s><s> Some	babies cry	more than others or for longer periods of time. </s><s> This is normal too. </s><s> It is also normal for a caregiver to become frustrated	Unclassifiable	High	Generic normal crying with no stated cause.	
41	nhs.uk	this to happen. </s><s> This can be hard for you, as it"s often the time when you"re most tired and least able to cope. </s><s> The amount	babies cry	tends to peak at about 7 weeks, then gradually tail off. </s><s> How to calm a crying baby </s><s> Try some of the following ways to comfort	Unclassifiable	High	General developmental pattern of crying and soothing; no specific cause.	
42	nhs.uk	and singing can keep your baby awake. </s><s> You might find lying them down after a feed will help. </s><s> Crying during feeds </s><s> Some	babies cry	and seem unsettled around the time of a feed. </s><s> If you"re breastfeeding, you may find that improving your baby"s	Discomfort	Medium	Crying/unsettled behavior occurs during feeding, suggesting feeding-related discomfort.	Fussiness
43	nhs.uk	during feeds can sometimes be a symptom of reflux, a common condition where babies bring back milk after feeds. </s><s> If your	baby cries	constantly </s><s> There are several reasons that can cause a baby to cry excessively. </s><s> It can be exhausting if you have tried	Unclassifiable	Medium	Reflux is mentioned nearby, but the KWIC introduces constant crying with several possible causes.	
44	nhs.uk	for your call. </s><s> If you decide to talk to your health visitor or GP, it can help to keep a record of how often and when your	baby cries	. </s><s> For example, this might be after every feed or during the evening. </s><s> This can help your health visitor or GP to work out if	Unclassifiable	High	Examples of when a baby cries are given to aid diagnosis, but cause is unresolved.	
45	nhs.uk	to check your baby"s positioning and attachment. </s><s> Sometimes you"ll notice your milk letting down in response to your	baby crying	or when you have a warm bath or shower. </s><s> This is normal. </s><s> How often should I feed my baby? </s><s> In the first week, your baby may want	Unclassifiable	High	Baby crying is only the stimulus for maternal milk let-down; cause unknown.	
46	hse.ie	them with you </s><s> Comforting your baby </s><s> Looking after yourself </s><s> Hearing your baby crying can be very upsetting. </s><s> Some	babies cry	more than others. </s><s> It is normal to feel stressed when your baby is crying, especially if nothing you do seems to be helping	Unclassifiable	High	General statement about caregiver stress in response to crying.	
47	hse.ie	your crossed arms. </s><s> Their legs and arms should gently fall either side of your arms </s><s> respond as quickly as possible to your	baby crying	. </s><s> Provide comfort by feeding, holding, rocking, singing or other method. </s><s> Leaving a baby to cry will cause more air to be	Unclassifiable	High	Soothing methods are listed, but the cause of crying is not specified.	
48	nhs.uk	usually starts when a baby is a few weeks old – and stops when they"re around 4 to 6 months of age. </s><s> If your otherwise healthy	baby cries	inconsolably for 3 or more hours a day, at least 3 days per week (and it"s been happening for 3 weeks or more) – it could be	Colic	High	Explicit diagnostic criterion for colic.	
49	nhs.uk	record of when your baby feeds, with details of how often and how much your baby brings the food back up, and how often your	baby cries	or seems distressed. </s><s> This will help your health visitor or GP decide if your baby needs treatment. </s><s> Gen. Anthony J.	Discomfort	High	Crying/distress is discussed in the context of reflux after feeds.	Unknown
50	nhs.uk	to feed. </s><s> If your baby is upset, try soothing them before feeding – a cuddle and some skin-to-skin contact may help. </s><s>	Babies cry	for lots of different reasons. </s><s> If they have been fed recently then hunger is unlikely to be the cause of their crying. </s><s> The	Multiple	High	Explicitly says babies cry for many reasons and discusses hunger as one possibility.	
51	nhs.uk	of new mums </s><s> Daily breastfeeding tips – on things like comforting your newborn, feeding frequency, and what to do if your	baby cries	during feeds </s><s> 3: TALK – Ask Breastfeeding Friend a specific question, for example "Hey Google, is my baby getting enough	Discomfort	Medium	Crying during feeds in a breastfeeding-support context.	
52	nhs.uk	? </s><s> If you"re already in a conversation with Breastfeeding Friend, you can just ask a question, such as "Why is my	baby crying	?" </s><s> If you"re not already in a conversation with Breastfeeding Friend, start by saying "OK/Hey Google, ask	Unclassifiable	High	Open-ended question asking why the baby is crying.	
53	nhs.uk	Friend? </s><s> If you"re in the Breastfeeding Friend just say "Alexa . . ." then ask the question e.g. "Alexa, why is my	baby crying	?" </s><s> In the general Alexa environment you need to use "Breastfeeding Friend" to open the Breastfeeding Friend e.g. "	Unclassifiable	High	Open-ended question asking why the baby is crying.	
54	eqi.org	murdered was this:-She had been set that night to mind Mrs. Hicks"s baby, and during the night she fell asleep, and the	baby cried	. </s><s> She, having lost her rest for several nights previous, did not hear the crying. </s><s> They were both in the room with Mrs.	Unclassifiable	High	Baby cries while caregiver sleeps; cause not given.	
55	lasc.us	the campfire did not give out its normal glow; the light acting like it was having a hard time penetrating the darkness. </s><s>	Babies cried	, dogs whimpered and young men found themselves shaking. </s><s> The unknown was stalking the land. </s><s> The thick darkness	Non-target	Medium	Fictional/atmospheric scene; no recoverable infant cry cause.	
56	ctk.org	, bold, and loving wife, Isabel. </s><s> Years later, after two miscarriages and one stillbirth, the grieving Isabel hears a	baby"s cries	on the wind. </s><s> A boat has washed up onshore carrying a dead man and a living baby. </s><s> DIGGING TO AMERICA by Anne Tyler In "	Non-target	High	Literary plot summary involving a found baby; no target cry cause.	
57	edc.org	looking at a healthy caregiver, their neural circuitry is wired to respond to successful caregiving behaviors. </s><s> The	baby cries	, the caregiver responds, the baby stops crying, and the caregiver"s brain releases chemicals that bring pleasure. </s><s> In	Unclassifiable	High	Generic caregiver-response sequence; no cause of cry.	
58	ntd.com	told the Stockton Record he was returning home from visiting a friend when he heard what sounded like a kitten and a	baby crying	at the same time. </s><s> Something told him to investigate, and it was then that he saw something moving inside the dumpster. </s>	Unclassifiable	Medium	Baby is found in a dumpster after crying is heard, but the immediate cause is not stated.	
59	pmq.com	young mother was ready to push while the pizzeria employees stood around her, holding up a screen for privacy. </s><s> "When the	baby cried	, I was cheering," Yarak recalled. </s><s> "[The woman] did an amazing job. </s><s> She was very brave." </s><s> During the labor process,	Birth	High	Baby cries immediately after delivery.	

Continued on next page

Table 9 – continued from previous page

Line	Reference	Left	KWIC	Right	Classification	Conf.	Contextual basis	Validation
60	ntd.com	County Sheriff"s Office deputies late June 6. </s><s> They responded after receiving a 911 call from someone who heard the	baby crying	. </s><s> Forsyth Sheriff Ron Freeman said at the time that the discovery of the baby was a miracle. </s><s> "It was, without doubt,	Unclassifiable	Medium	Abandoned baby is heard crying; cause is not stated.	Abandoned baby
61	ntd.com	. </s><s> About five minutes after the call, the caller told emergency services that the baby was breathing. </s><s> "That"s the	baby crying	. </s><s> We"re getting vitals from mom right now," the caller said. </s><s> Phoenix police have launched a sexual assault	Birth	Medium	Emergency context strongly suggests a just-delivered baby; crying confirms breathing.	
62	esf.edu	. </s><s> Hearing Ear: Alerts persons with hearing impairments to specific sounds within the environment, such as a	baby crying	, a telephone ringing, a fire alarm, or door bell; </s><s> Fetch & Carry: Assists mobility impaired persons by obtaining and	Non-target	High	Service-animal/accessibility example.	
63	apl.org	she is on a collision course to meet them. </s><s> Twenty-five years ago, police were called to 16 Cheyne Walk with reports of a	baby crying	. </s><s> When they arrived, they found a healthy ten-month-old happily cooing in her crib in the bedroom. </s><s> Downstairs in the	Unclassifiable	High	Healthy infant is heard crying in crib; no cause supplied.	
64	edj.net	<s> My family and in-laws came to the house here during that time. </s><s> We had just finished building it. everyone was scared. </s><s> The	babies cried	. </s><s> There were about twenty people here for 15 days." </s><s> Don Rafael"s house was indeed small. </s><s> Made of brick, adobe and	Unclassifiable	Medium	Babies cry during a frightening/disaster situation, but cause is not specific enough.	
65	bpr.org	there are fireworks. </s><s> It plays me fan sounds through my wife"s snore. </s><s> It always soothes me with gentle birdies. </s><s> When my	baby cries	, I ignore. </s><s> MORENO: That is a white noise machine. </s><s> COULTON: White noise machine - that is correct. </s><s> EISENBERG: I do find it	Non-target	High	Comedic song/quiz line about a white-noise machine.	
66	cin.org	, defenceless, even to the point of lacking that minimal form of defence consisting in the poignant power of a newborn	baby"s cries	and tears. </s><s> The unborn child is totally entrusted to the protection and care of the woman carrying him or her in the womb. </s>	Non-target	High	Rhetorical comparison involving newborn cries, not a crying episode.	
67	tchd.us	. </s><s> your baby"s fever is higher than 102o when taken in the armpit, 45 minutes after the last dose of acetaminophen.your	baby cries	or screams for 3 or more hours after the DTaP shot and can"t be comforted. </s><s> your baby has a high pitched unusual cry within	Vaccination	High	Prolonged inconsolable crying is explicitly post-DTaP shot.	
68	cbr.com	Mark Pacella"s Alpha Flight #106 started with the team fighting Mr. Hyde in the streets of Toronto. </s><s> Northstar heard a	baby crying	from inside of a dumpster. </s><s> Pulling the newborn out, he abandoned the fight and flew to the nearest hospital. </s><s> After an	Non-target	High	Comic-book fictional scene.	Abandoned baby
69	uhcg.us	. </s><s> Even the jackals feed their young, but not my people, Israel. </s><s> They are like cruel desert ostriches, Heedless of their	babies" cries	. </s><s> The children"s tongues stick to the roofs of their mouths for thirst, for there is not A drop of water left. </s><s> Babies cry	Hunger	High	The surrounding passage explicitly describes thirst/deprivation.	
70	uhcg.us	cries. </s><s> The children"s tongues stick to the roofs of their mouths for thirst, for there is not A drop of water left. </s><s>	Babies cry	for bread, but no one can give them any. </s><s> Those who used to eat fastidiously are begging in the streets for anything at all. </s>	Hunger	High	Explicit: babies cry for bread.	
71	jwa.org	other way. </s><s> These images are impossible to make real without the sacrifice of a partner. </s><s> Someone has to step out when the	baby"s cries	become overly disruptive or to help set up kiddush when the congregation is in need of a couple hands. </s><s> When a woman takes	Unclassifiable	High	Realistic generic crying in synagogue; cause unknown.	
72	zop.org	in the age of 2 to 4 months. </s><s> The infants die during sleep and without showing any sign that they have suffered, or that the	baby cried	when it happened. </s><s> There is no treatment or even a clear specific prevention of SIDS, even with years of medical research	Non-target	High	SIDS passage explicitly says there may have been no crying.	
73	iapm.ca	if the Germans were close by. </s><s> His father disagreed, and felt they should not turn these people away. </s><s> As feared, the	baby"s cries	did alert the Germans and everyone was marched out of the cellar and put aboard trains. </s><s> While being transported,	Unclassifiable	High	Baby cries while people hide from soldiers; cause of cry is not stated.	
74	sti.edu	these options doesn"t guarantee that the video conferencing tools aren"t collecting data on your surroundings. </s><s> If a	baby cries	in the background, you might see ads for pacifiers soon. </s><s> If your pet pops up on screen, ads for pet supplies might show up	Non-target	High	Hypothetical advertising/privacy example.	
75	cbr.com	name for God. </s><s> More visions are displayed on the page, with a massive face taking the majority of the focus away from a	baby crying	and a city skyscraper crumbling to the ground. </s><s> Psylocke is told to go to Tokyo, with another clue stating that Apoth has	Non-target	High	Comic-book imagery.	
76	bpr.org	conferencing platform. </s><s> But you have to admit that nothing says quarantine quite like hearing a stray turn signal or a	baby crying	in the background of your work meeting or virtual family gathering. </s><s> Comic Nore Davis has heard it all because when New	Non-target	High	Generic background-noise example in remote meetings.	
77	bpr.org	- says she has one word for this experience. </s><s> AIZENMAN: Nights are the worst. </s><s> BRITT: You have a bunch of people talking,	babies crying	. </s><s> They keep it so cold. </s><s> I"ve got a sheet, two blankets and a quilt on my bed, and I got a real thick housecoat. </s><s> And I still	Unclassifiable	Medium	Babies are crying in a crowded/cold setting, but the line does not establish why.	Sleepiness
78	jnj.com	bed after a long day, placing your head on a soft pillow, and drifting off to sleep–only to be abruptly awakened by your	baby"s cries	. </s><s> For many moms and dads, this repeat scenario can go on for months, and in some cases, even years. </s><s> But getting enough	Unclassifiable	High	Night waking with a baby's cries; cause is not specified.	
79	bpr.org	been there. </s><s> A couple hours later, they got another 911 call – the baby had been found. </s><s> Griffin said the woman who heard the	baby"s cries	summoned her husband, who climbed down into the 75-foot ravine and found the little girl. </s><s> Once Madden was questioned at	Unclassifiable	Medium	Abandoned baby is found after cries are heard; precise cause is not stated.	
80	life.ca	hard either to go to sleep or wake up. </s><s> Could these social facts be connected? </s><s> I strongly suspect they are." </s><s> ~ John Holt </s><s> "A	baby"s cry	is precisely as serious as it sounds." </s><s> ~ Jean Liedloff </s><s> "A newborn baby has only three demands. </s><s> They are warmth in the arms	Multiple	Medium	The adjacent quotation lists warmth, milk and security as newborn needs associated with crying.	
81	cbr.com	him that their connection will transcend the generations, however, and as the screen fades out, we hear a newborn	baby"s cry	… </s><s> 1 Episode 39: "Venom of the Red Lotus" (9.6) </s><s> This is the highest-rated The Legend of Korra episode of all, and it"s	Non-target	High	Television episode sound cue.	
82	wvi.org	how I tried to give the baby water to keep her calm. </s><s> But I thank God I didn"t." </s><s> Baindu could hardly endure the scene of her	baby crying	and was on the verge of giving up the notion of exclusive breastfeeding when a member of the "Mamaneiva" (Grandmother)	Unclassifiable	Medium	The mother considers giving water because the baby is crying, but the text does not establish hunger/thirst as the cause.	
83	pets.ca	hers. </s><s> Even cats that have a high percentage of siamese blood are LOUD, actual siamese are even louder. </s><s> They sound like a	baby crying	, and this will go on and on if you"re not giving them enough attention, and for some cats they just like to talk. </s><s> I don"t	Non-target	High	Cat vocalization compared with a baby crying.	
84	pets.ca	attention, and for some cats they just like to talk. </s><s> I don"t think some people understand what I mean by sounds like a	baby crying	.... </s><s> I mean sounds like a baby crying....just as loud just as annoying if you can"t handle it. </s><s> They are very smart cats,	Non-target	High	Cat vocalization compared with a baby crying.	
85	pets.ca	just like to talk. </s><s> I don"t think some people understand what I mean by sounds like a baby crying.... </s><s> I mean sounds like a	baby crying	....just as loud just as annoying if you can"t handle it. </s><s> They are very smart cats, and they do tend to be dog like, but they	Non-target	High	Cat vocalization compared with a baby crying.	
86	npaf.ca	pills aren"t the only problem. </s><s> They"re also seeing an increase in meth use and other street drugs. </s><s> The reality is, all	babies cry	: It"s the best (and only) way for them to communicate their needs at this tender age. </s><s> And as parents, we"re biologically	Unclassifiable	High	Generic statement that crying communicates infant needs; no single cause.	
87	jjg.net	on this one because I"m packing a lot of different factors up into this one category. </s><s> There"s the context of the moment:	babies crying	, dogs barking, phones ringing. </s><s> (Calgon, take me away!) Then there"s personal context: the history, associations,	Non-target	High	Generic list of distracting background noises.	
88	voy.com	your husband," the doctor said laughing. </s><s> "PUSH Marina" the doctor said, and Marina said, the next thing we heard was a	baby"s cry	. </s><s> "It"s a girl," the doctor said. </s><s> "It"s a girl, Imanol, " Marina said. </s><s> "She is beautiful, just like her mother," Imanol	Birth	High	Baby cries immediately after the mother pushes during delivery.	
89	asac.us	containing oxytocin or a water placebo. </s><s> During the scan, the mothers will view pictures of their baby and hear their	baby"s cry	. </s><s> This involves a mother interview, videotape play and 2 brain scanning sessions. </s><s> There are a total of four visits over a	Non-target	High	Recorded baby cries are experimental stimuli.	
90	voy.com	Thread	Next Message	Previous Message] Date Posted: 12/14/06 8:45pm In reply to: Nikki "s message, "letting	baby cry	? and other questions..." </s><s> on 12/ 6/06 12:18pm I just wanted to say thank you for all your wonderful advice. </s><s> I have been	Unclassifiable	Medium	Forum thread concerns letting a baby cry, but this concordance line does not establish the cause.	
91	voy.com	working out all those quirks in the first few months. </s><s> I think the advice given was excellent and the little bit your	baby cries	is perfectly normal. </s><s> Your "baby"s first" with Santa is only the first of many sweet and exciting firsts of motherhood; I	Unclassifiable	High	Normal baby crying is mentioned, but no cause is identified.	
92	voy.com	Thread	Next Message	Previous Message] Date Posted: 01/12/07 5:42pm In reply to: Nikki "s message, "letting	baby cry	? and other questions..." </s><s> on 12/ 6/06 12:18pm I work in the PC field and I read an article in the past that I saved. </s><s> It was	Unclassifiable	Medium	Forum thread concerns letting a baby cry, but this concordance line alone does not establish a sleep-related cause.	

Continued on next page

Table 9 – continued from previous page

Line	Reference	Left	KWIC	Right	Classification	Conf.	Contextual basis	Validation
93	voy.com	Thread	Next Message	Previous Message] Date Posted: 12/11/06 6:31pm In reply to: Nikki ”s message, "letting	baby cry	? and other questions..." </s><s> on 12/ 6/06 12:18pm Hi.. </s><s> I think Luka”s advice about sleeping is pretty sound (no pun intended	Sleepiness	High	Reply explicitly discusses sleeping in the ’letting baby cry’ thread.	
94	voy.com	they aborted. </s><s> Their life is filled with regret and shame. </s><s> They wake up in the middle of the night convinced they heard a	baby crying	. </s><s> They are unable to look at a pregnant woman the same anymore. </s><s> These are not threats or ways of trying to get you to think	Non-target	High	Imagined cry in a description of post-abortion distress.	
95	voy.com	Previous Thread	Next Message	Previous Message] Date Posted: 12/ 6/06 9:44pm In reply to: Nikki ”s message, "letting	baby cry	? and other questions..." </s><s> on 12/ 6/06 12:18pm Are you breast feeding or bottle feeding? </s><s> If breast feeding you might find	Hunger	Medium	Reply to crying thread immediately asks about breast- versus bottle-feeding, making feeding/hunger the salient hypothesis.	Unknown
96	bgmi.us	of another it may the real evocation of a similar feeling in themselves in response to another. </s><s> (Empathy). </s><s> When the	baby cries	the healthy mother/father responds with empathy. </s><s> They literally feel distress - but not for themselves rather for the	Unclassifiable	High	Generic empathy example; cause of infant crying unspecified.	
97	evo.org	Bandana. </s><s> Released in the US by Sire on CD (40135-2) in a digi-pak featuring a blow-up of the artwork of BAD1015. </s><s> "Cry	Baby Cry	" written by Lennon and McCartney. </s><s> Mat, a reckless drifter pushing 40, shows up unexpectedly at his gainfully employed	Non-target	High	Song title/lyric reference.	
98	voy.com	a bunch of men trying to gang-rape me, but at the last minute, I got away. </s><s> Women have told me of nightmares where they hear	babies crying	, and they can”t reach them. </s><s> Stuff like that. </s><s> Abortion has a huge emotional impact on a woman. </s><s> You may be able to make	Non-target	High	Nightmare about hearing babies crying.	
99	aop.com	? </s><s> Last night”s dishes, last week”s laundry, and yesterday”s schoolwork all clamor for your attention. </s><s> Your	baby”s cry	reminds you of yet another need, and your young toddlers seek your affection and time for play. </s><s> On top of that, your	Unclassifiable	High	Actual household baby cry, but the need causing it is not specified.	
100	tcj.com	this was a mortal enemy. </s><s> And I really do believe that infants know just whom to trust. </s><s> Some instinct tells them. </s><s> We know	babies cry	terribly when they”re held wrong, because some instinct says, "This jerk is going to drop me," or "My neck is going to	Discomfort	High	Crying is attributed to being held incorrectly/feeling physically insecure.	

Whose Robots.txt? Consent-Infrastructure Inequality Across the Languages of Web Corpora

Austin Cheng* **Gary Zhang***
William A. Shine Great Neck South High School
341 Lakeville Road, Great Neck, NY 11020
austincheng98@gmail.com zhanggary607@gmail.com
* Equal contribution.

Abstract

Website owners increasingly restrict AI crawlers through robots.txt. Audits are English-centric and do not stratify results by language. We measure opt-out for 4,900 domains across fourteen languages and three resource tiers in FineWeb and FineWeb2. Among reachable domains, high-tier sites block 28.8 percent, versus 11.2 percent and 9.8 percent in the mid and low tiers; an exact language-partition permutation gives p = 0.008 for the high-low gap. The gradient is concentrated in naming: files naming any of fifteen AI agents fall from 33.6 percent to 10.6 percent across tiers, while named-agent blocking shows no tier ordering. Among files listing conventional crawlers, AI naming falls from 64.0 percent to 19.0 percent, consistent with vocabulary diffusion rather than maintenance failure. Applying current robots.txt retroactively to an earlier crawl, design-weighted document mass behind a Common Crawl opt-out is 31.2 percent for English; tier means are 26.5 percent, 8.9 percent, and 10.7 percent. Reachability falls from 78 percent in English to 39 percent in Uyghur. These results show a gradient in recorded policy that a consent-respecting rebuild would inherit. We release every fetched robots.txt, derived rates, and code.[1]

1 Introduction

The web is closing to AI crawlers: Longpre et al. (2024) found that 28% of the most actively maintained sources behind C4, RefinedWeb, and Dolma now restrict AI agents through robots.txt, up from near-zero twelve months earlier. Corpus builders honor these signals, making robots.txt an operational consent layer for pretraining data.

Prior audits do not stratify by language. Writing an AI clause into robots.txt requires knowing that GPTBot exists, maintaining a server configuration, and following a policy debate conducted in English. If those preconditions concentrate in wealthy, English-speaking web infrastructure, recorded opt-out should fall with the resource level of the language a domain serves, so filtering by robots.txt may remove recorded refusals rather than preferences.

We test this prediction using FineWeb (Penedo et al., 2024) and FineWeb2 (Penedo et al., 2025), which ship source URLs for every document. We sample 350 domains per language in two stages, fetch each live robots.txt, and parse fifteen AI crawlers. Domain-level rates are primary; composition estimates carry a weaker Horvitz–Thompson correction. The question is whether the gradient is in naming an AI agent, survives size and site-type controls, and changes the composition a consent-respecting rebuild would inherit.

2 Related Work

Longpre et al. (2024) established the longitudinal, English-centric picture by auditing 14,000 domains behind three corpora and documenting the rapid growth of AI-specific restrictions, along with inconsistencies between robots.txt and terms of service. Their results contain no per-language or per-region stratification of consent, which is the axis we add. Dodge et al. (2021) set a template for auditing what corpus construction silently does to composition; we apply that question to the consent layer. Stratified robots.txt audits along other axes find a similar selection effect: reputable news sites block AI crawlers at much higher rates than misinformation sites (Steinacker-Olsztyn et al., 2026), suggesting that consent filtering can skew corpora along a quality axis. At larger scale, Cui et al. (2025) audit 610,681 Tranco-listed domains for eighteen LLM bots, surfacing vendor-inconsistent rules and outright violations, but report no per-

[1] https://github.com/austin-cheng98/robots-txt

Proceedings of the 13th Web-as-Corpus Workshop (WaC-13), pages 347–355
October 29, 2026 ©2026 Association for Computational Linguistics

language breakdown. The robots exclusion protocol on which these audits rely began as an informal convention between crawler operators and site owners (Koster, 1994).

On the corpus side, FineWeb2 and its predecessor publish per-document URLs, which we use as the sampling frame (Penedo et al., 2024, 2025). Multilingual data-quality work has focused on what pipelines admit, including wrong-language contamination and coverage gaps (Kreutzer et al., 2022; Caswell et al., 2020); we ask instead who is in a position to keep text out. This shifts the unit of concern from the pipeline alone to the population feeding it: a consent-respecting filter may remove text unevenly even when the corpus builder applies one rule consistently across languages.

3 Method

3.1 Sampling

We take fourteen language subsets across three resource tiers: English, German, Japanese, and French (high); Indonesian, Turkish, Vietnamese, Thai, and Azerbaijani (mid); Swahili, Sundanese, Yoruba, Uyghur, and Irish (low). Tiers track Joshi et al. (2020): every high-tier language is class 5, every low-tier language classes 1–2, and the mid tier classes 3–4 apart from Azerbaijani (class 1), whose placement we test in Section 4. English comes from FineWeb `sample-10BT` and the other thirteen from FineWeb2, which has no English subset; both derive from Common Crawl through near-identical pipelines, but English's census is nonetheless drawn from a differently filtered corpus. We stream up to 200,000 documents per subset (the cap binds everywhere except Sundanese, Yoruba, and Uyghur) and aggregate document and token counts by domain. Our domain unit is the URL hostname with a leading `www.` stripped, not the registered domain or eTLD+1, so a publisher network spread over regional subdomains contributes several units. Frame membership comes from the pipeline's own language identification, so multilingual hosts appear under more than one language and language-ID error enters our frame (Caswell et al., 2020).

The census retains the top 20,000 domains by document mass and discards the rest before sampling. The cap binds for all four high-tier languages, four of the five mid-tier ones, and no low-tier one, discarding 117,447 English and 82,783 Japanese domains, so every estimate below is for the top-20,000-domain web of a language, a population systematically narrower for the high tier. Within it we draw 350 domains in two stages, the top 175 by document mass with certainty and 175 uniformly at random from the remaining $N - 175$. The tail weight varies six-fold, from 18.7 for Uyghur to 113.3 for English, so simply adding the mass of sampled domains, as an earlier version did, gives every head domain the weight of a tail domain and describes the sampled head rather than the language. Every mass-weighted quantity below is instead a Horvitz–Thompson estimate (Horvitz and Thompson, 1952), defined in the caption of Table 2. Domain-level rates carry no weights, and the two stages double as a size control.

3.2 Parsing

We fetch `https://domain/robots.txt` with an 8-second timeout, falling back to HTTP, and archive every response body. We parse user-agent blocks for fifteen AI-associated crawlers (GPTBot, CCBot, ClaudeBot, Google-Extended, Bytespider, and ten more), recording a full opt-out when an agent receives `Disallow: /`; a domain counts as opted out when it fully blocks at least one. We also record whether a file names any of the fifteen in any rule including `Allow` rules, which factors opt-out into naming an AI crawler and blocking a named one. Relabeling a stratified sample of 60 archived bodies by hand reproduces the parser verdict in 58, an agreement of 96.7%, the two disagreements running in opposite directions: `nli.ie` gives Bytespider a `Disallow: *` rather than a literal slash, which our rule excludes and the manual label counts as a block, and `pleinelune.niceboard.com` has a body truncated at 6,000 characters inside an old-style scraper blocklist, so the parser records a block the visible text cannot support. Domain-level rates carry 2,000-draw bootstrap intervals.

3.3 Policy

Both corpora derive from Common Crawl, so the restriction a rebuild would be bound by is the one addressed to CCBot, which we report as the rebuild policy alongside the fifteen-agent union as the broader "any AI agent" reading, GPTBot-only, and two variants folding in a blanket `User-agent: *` with `Disallow: /` (Table 3). Since tier is assigned at the language level, domains within a language are not independent un-

Language	Reach	Names	Opt-out [95% CI]	Mass	n	Blk
High tier						
English	.78	.493	.467 [.408, .529]	.392	134	.95
German	.73	.402	.299 [.244, .354]	.426	102	.75
French	.77	.264	.230 [.182, .279]	.271	71	.87
Japanese	.65	.159	.128 [.088, .173]	.143	36	.81
Mid tier						
Indonesian	.64	.276	.138 [.093, .187]	.084	62	.50
Thai	.61	.168	.126 [.084, .173]	.163	36	.75
Turkish	.68	.214	.109 [.071, .147]	.088	51	.51
Azerbaijani	.63	.131	.104 [.068, .144]	.095	29	.79
Vietnamese	.61	.089	.080 [.047, .113]	.062	19†	.89
Low tier						
Sundanese	.47	.129	.123 [.074, .178]	.077	21	.95
Yoruba	.43	.113	.113 [.066, .166]	.154	17†	1.00
Uyghur	.39	.109	.109 [.058, .167]	.123	15†	1.00
Irish	.58	.124	.104 [.064, .149]	.123	25	.84
Swahili	.58	.064	.054 [.025, .089]	.094	13†	.85

Table 1: AI-crawler opt-out by language. Reach is the share of 350 sampled domains answering robots.txt; Names and Opt-out are, among reachable domains, the shares naming at least one of fifteen AI agents and fully blocking one, respectively; bracketed intervals are bootstrap 95% CIs. Mass is design-weighted document mass behind a fifteen-agent block under today's robots.txt applied retroactively; n counts reachable domains naming an agent and Blk is the share of them blocking one. † marks $n < 20$, where Blk is a count.

der a tier-shuffle null, so our primary test is an exact permutation over the language partition (126 high-versus-low splits), equal-weighted unless we say otherwise; domain-pooled shuffles overstate precision and serve as descriptives only.

Live robots.txt records policy today while the corpora were crawled earlier, so every composition number below is what today's robots.txt would do to an earlier crawl and not a forecast of a future crawl, the frame of Longpre et al. (2024). Unreachable domains are reported rather than dropped, since reachability itself tracks the infrastructure gap under study.

4 Results

4.1 Consent

English sits alone at the top of Table 1 with 46.7% of reachable domains blocking at least one AI crawler, German next at 29.9%. Pooled within tiers, reachable high-tier domains block at 28.8% ($n = 1{,}021$) against 11.2% ($n = 1{,}112$) and 9.8% ($n = 857$), a 19.0-point high–low gap with cluster permutation $p = 0.008$ equal-weighted and 0.032 count-weighted, and the English CI $[.408, .529]$ overlaps no mid- or low-tier interval. Taking each language as one unit the tier means are 28.1%, 11.1%, and 10.0%, and testing the high tier against all ten remaining languages widens the reference set to 1,001 splits and sharpens the verdict to $p = 0.002$. The step sits at the high tier, since mid and low are not separable ($p \geq 0.45$), so this is a two-level split rather than a three-way ordering, and tier rank predicts opt-out at $\rho = 0.69$ ($p = 0.009$). None of this is mass-weighted or changes under the design weights below.

Figure 1 summarizes the decomposition. Opt-out factors into whether a robots.txt names any of the fifteen agents and whether a named agent is fully blocked, and only the first step shows the gradient. Pooled naming runs 33.6%, 17.7%, and 10.6%, a 23.0-point high–low gap with cluster $p = 0.008$ ($p = 0.007$ against all ten remaining languages, $\rho = 0.77$, $p = 0.002$), and English alone reaches 49.3%. CCBot (428 blocks), GPTBot (422), and ClaudeBot (399) lead in every tier (rank correlations 0.89 to 0.94). Conditional on naming, full blocking shows no tier ordering we can detect: 85.7% (high, $n = 343$), 62.9% (mid, $n = 197$), and 92.3% (low, $n = 91$), cluster $p = 0.167$, bootstrap 95% interval on the pooled gap $[0.6, 13.0]$ points. That is consistent with no ordering rather than evidence of its absence, and the denominators say why: four languages have fewer than twenty naming domains, the 100.0% cells for Yoruba (17 of 17) and Uyghur (15 of 15) are counts not rates, and the low-tier 92.3% rests on 91 domains.

4.2 Mechanisms

The mid tier's 62.9% comes almost entirely from Indonesian and Turkish sites, without which it is 79.8%, and in both one template rather than a distribution of site-level decisions does the work. Of the 31 Indonesian domains naming an agent without blocking it, 28 are regional subdomains of `tribunnews.com`, whose shared robots.txt places Google-Extended and GPTBot in the same `Allow: /` group as Googlebot. Of the 25 Turkish ones, 20 share a news-CMS template that stacks eight AI user-agent lines and then opens a `User-agent: *` group of path-prefix disallows, so the named agents receive partial rules rather than `Disallow: /`, exactly where recorded policy and preference come apart.

Three mechanisms behind a gap in recorded policy are partly observable. If low-tier operators simply maintain their files less, they should name conventional crawlers less too, and they do not: naming at least one of fifteen conventional crawlers (Googlebot, Bingbot, AhrefsBot, and others) runs

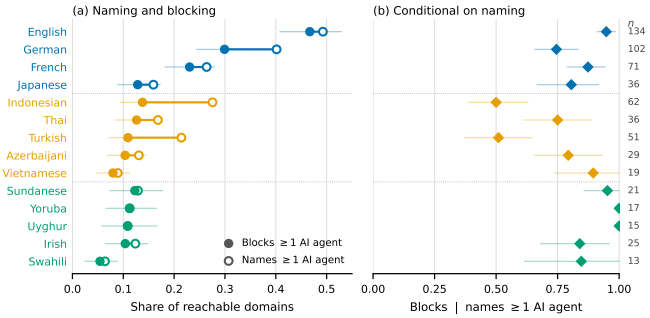


Figure 1: Consent gradient and decomposition. Filled/open markers show blocking/naming rates; panel (b) conditions blocking on naming and labels the naming count n. Colors mark tiers.

35.7%, 19.0%, and 24.6%, neither monotone nor significant (gap 9.7 points, cluster $p = 0.190$), while AI naming over the same files has $p = 0.008$. Restricting to files that already curate a conventional list, the share also naming an AI agent is 64.0% (high, $n = 364$), 44.5% (mid, $n = 211$), and 19.0% (low, $n = 211$), a 39.7-point gap with cluster $p = 0.008$. A maintenance-budget account predicts both rates fall together and they do not, while diffusion of the AI-agent token set predicts exactly this. The "aware and permitting" signal agrees, since files naming an AI agent and blocking none are 4.8%, 6.6%, and 0.8% of reachable domains ($p = 0.008$), essentially absent in the low tier. Two caveats belong here. Low-tier conventional naming is partly inherited, since SEO blocklists for SemrushBot, MJ12bot, and AhrefsBot ship with some templates and dominate the low tier's non-Google entries. And stock CMS defaults are real: an untouched WordPress, Blogger, or Wix body is 25.2% of the low-tier sample against 9.0% of the high-tier one. Restricting to files somebody edited, the gradient survives at 33.0%, 14.1%, and 13.1% (cluster $p = 0.016$), so platform defaults are part of the gap and not the whole of it.

Site-type controls point the same way without settling much, since the gradient holds within ccTLDs (31.7%, $n = 432$; 12.2%, 329; 13.3%, 181) and within gTLDs (28.4%, 510; 10.7%, 707; 9.8%, 590), and platform-hosted subdomains, 12.0% of the low-tier sample against 2.1% of the high-tier one, almost never opt out, yet removing them moves the pooled gap only from 19.0 to 18.4 points, the low tier's non-platform rate of 11.0% still far below the high tier's 29.4%. Commercial incentive, legal exposure, and publisher type remain live alternatives our data cannot separate; Table 4 carries the remaining splits.

4.3 Composition

Under today's robots.txt applied retroactively to a corpus crawled earlier, and taking the CCBot restriction a rebuild would be bound by, design-weighted document mass behind a block is 31.2% for English, with tier means of 26.5%, 8.9%, and 10.7% (cluster $p = 0.032$ high–low, 0.005 high-versus-rest); the CCBot rate at the domain level is 23.8%, 9.6%, and 8.5% ($p = 0.032$ and 0.007). Reading "any AI agent" instead raises English to 39.2% of document mass (41.4% of tokens) and the tier means to 30.8%, 9.8%, and 11.4% on document mass, 27.6%, 8.7%, and 11.2% on tokens; Figure 2 compares these domain and mass estimands. Design weighting materially weakens this part of the paper: under the CCBot policy, as under any-15, the low tier sits above the mid tier on mass, so what survives is a high-tier step and not a three-way gradient; high–low under any-15 gives $p = 0.016$ on document mass but $p = 0.10$ on tokens, which is not significant, while high-versus-rest survives both ($p = 0.003$ and 0.015), and the rank correlation with tier falls to $\rho = 0.51$ ($p = 0.069$) and $\rho = 0.22$ ($p = 0.45$). The unweighted figures of an earlier version, 62.4% for English and tier means of 44.3%, 13.3%, and 12.2%, describe the sampled head: the top 175 English domains by mass were fetched with certainty, and 63.9% of the mass of those that answered sits behind a block. Table 2 gives per-language mass.

The effect is robust to which AI agent defines the policy and not to folding in a blanket `User-agent: *` with `Disallow: /`. Wildcard-only blocking runs slightly against the tier order (8.5%, 9.7%, and 10.9% pooled, $p = 0.37$), since low-tier sites use blanket blocks more

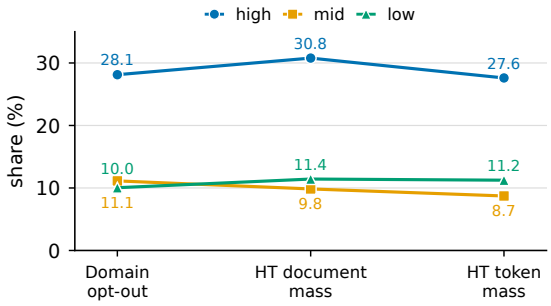


Figure 2: Tier means under three estimands. Domain opt-out is the reachable-domain rate; design-weighted document and token mass are Horvitz–Thompson estimates under the fifteen-agent policy.

than AI-specific clauses. Counting one as an opt-out alongside the fifteen agents gives 32.9%, 13.1%, and 14.7%, which still separates high from low ($p = 0.040$) but inverts mid and low, and combining it with the CCBot policy is the one variant where the high–low test fails, at $p = 0.064$, though high-versus-rest survives every variant at $p \leq 0.008$. A reader who takes a blanket block to express the same refusal as a named AI clause should read our claim as a high-tier step with no ordering below it.

4.4 Robustness

Azerbaijani is Joshi class 1 but sits in our mid tier, and neither correction disturbs anything: excluding it leaves the high–low test at $p = 0.008$, moving it to the low tier strengthens that test to $p = 0.005$, and the rank correlation rises in both cases. Treating a 404 or 410 as an observed policy of no restriction rather than a failed fetch cleans the gradient slightly, with tier means of 25.9%, 10.3%, and 8.0%, high–low $p = 0.008$, and $\rho = 0.76$ ($p = 0.003$).

High-tier samples contain larger publishers and large sites block more, so the gradient could be a size effect, which the two stages check directly. Among the 175 head domains per language, high-tier sites block at 40.4% ($n = 559$) against 10.6% ($n = 538$) in the low tier, cluster $p = 0.008$; among tail domains the rates are 14.7% ($n = 462$) and 8.5% ($n = 319$), pointing the same way but not distinguishable from chance ($p \geq 0.30$), so the tail is not confirmation. A head domain is large only relative to its own language, and the top 175 hold 10.9% of retained English document mass against 85.9% of retained Uyghur mass, so this is a within-language stratum control, not a matched comparison.

Between 65% and 78% of sampled high-tier domains answer a robots.txt request against 39–58% in the low tier (cluster $p = 0.008$, $\rho = 0.92$, $p < 10^{-4}$). DNS failure, the domain no longer resolving, rises from 8.9% of high-tier fetches to 19.9% and 26.3% and accounts for roughly half of mid- and low-tier failures, while 403 is flat at 4.6%, 4.6%, and 4.5%. The gradient is therefore more consistent with link rot than with a single cloud vantage point, though geographic blocking remains unmeasured. Per-language failure types are in Table 5.

5 Discussion

The decomposition and the proxies narrow the space of mechanisms without closing it. Capacity alone does not fit, since conventional-crawler naming shows no significant tier gradient while AI naming among the very files that curate a conventional list falls from 64.0% to 19.0%. Deliberate permission does not fit the low tier either, since a web that engaged with AI crawlers and chose to admit them would produce naming without blocking, and that pattern is 0.8% of low-tier files. Platform defaults are a component and not the whole, since restricting to edited files leaves a gradient at $p = 0.016$. Awareness and infrastructure is the interpretation the evidence best supports, and it remains an interpretation, since we cannot observe the preferences of domains that never mention an AI agent and silence equals neither refusal nor consent. What the data establish is narrower, that the operative consent mechanism transmits a gradient in recorded policy which any filter built on robots.txt inherits. Two designs would settle the rest: surveying low-tier operators, which measures awareness as no file-level proxy can, and a category-matched comparison sampling news sites against news sites across languages, of which our hostname-keyword and TLD splits are the weakest possible version.

Honoring robots.txt beats ignoring it, but treating it as the consent layer privileges the infrastructure-rich web. Corpus builders can act by reporting per-language consent-filtered composition alongside raw composition in every data card, which our release makes cheap for any corpus shipping source URLs.

6 Limitations

6.1 Sampling

We sample 350 domains per language in two stages from censuses capped at 20,000 domains, so every estimate generalizes to the top-20,000-domain web of a language rather than to its web, and that restriction bites hardest in exactly the languages with the most domains. Mass-weighted quantities depend on a head of 175 domains that a different draw would mostly re-select, and the design-weighted estimates inherit large tail weights, up to 113.3 for English, which makes them noisier than their point estimates suggest. Domain-level rates are the quantity our design supports best and we have made them the headline accordingly. English's census comes from FineWeb rather than FineWeb2, because FineWeb2 has no English subset; the pipelines are near-identical and both corpora derive from Common Crawl, but we cannot rule out that the difference contributes to English's position.

6.2 Measurement

All fetches issue from one cloud vantage point. The failure-type breakdown argues against this producing the gradient, since 403s are flat across tiers while DNS failures are not, but geo-dependent robots.txt and geo-dependent blocking go unmeasured and the released per-domain failure types let a reader audit this. Language attribution inherits whatever errors the corpus pipeline's language identification makes, and a domain serving several languages can enter more than one frame. Our domain unit is a hostname, so a publisher network spread over regional subdomains contributes many units, which matters most for the conditional-blocking analysis.

Robots.txt is the mechanism corpus builders honor, but preferences can also be recorded in the TDM Reservation Protocol (W3C Text and Data Mining Reservation Protocol Community Group, 2024), in `ai.txt` proposals and `noai` values in `robots` meta tags and `X-Robots-Tag` headers, and in terms of service no crawler parses (Longpre et al., 2024). All demand the same infrastructure literacy and vendor vocabulary, so we expect them to track the gradient rather than offset it, while CDN-managed AI blocking plausibly cuts the other way.

Our fifteen-agent roster follows current industry practice, but sites can also express preferences in terms of service or use emerging standards we do not parse, and rates under the named-agent criterion are lower bounds on expressed restriction. Expressed restriction is in turn not enforcement, since crawlers comply with robots.txt selectively (Kim et al., 2025; Liu et al., 2025), so what we measure is the demand to be excluded rather than the supply of compliance. The retroactivity caveat applies throughout: 2026 robots.txt cannot bind crawls that already happened.

6.3 Scope

Fourteen languages are not the web and our tiers are coarse. The gradient we report is a statement about the fourteen subsets we sampled, since the language-cluster tests treat those fourteen as the population of interest rather than as a sample drawn from all languages, and nothing here licenses an extrapolation to the roughly 1,800 languages FineWeb2 covers.

Ethics Statement

Our measurement reads only files that site operators publish for automated clients, at one request per domain, and the release contains robots.txt bodies together with the domain names already present in the public FineWeb and FineWeb2 URL fields. We collect no personal data and no page content. Naming domains that do not block AI crawlers could in principle direct scrapers toward them, but those domains already appear in the public corpora we sample from, so the release adds no exposure the corpora do not already create.

References

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.

Jian Cui, Mingming Zha, XiaoFeng Wang, and Xiaojing Liao. 2025. The odyssey of robots.txt governance: Measuring convention implications of web bots in large language model services. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, pages 21–35, Taipei, Taiwan. Association for Computing Machinery.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

D. G. Horvitz and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Taein Kim, Karstan Bock, Claire Luo, Amanda Liswood, Chloe Poroslay, and Emily Wenger. 2025. Scrapers selectively respect robots.txt directives: Evidence from a large-scale empirical study. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25)*, pages 541–557, Madison, WI, USA. Association for Computing Machinery.

Martijn Koster. 1994. A standard for robot exclusion. `https://www.robotstxt.org/orig.html`.

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.

Enze Liu, Elisa Luo, Shawn Shan, Geoffrey M. Voelker, Ben Y. Zhao, and Stefan Savage. 2025. Somesite I used to crawl: Awareness, agency and efficacy in protecting content creators from AI crawlers. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25)*, pages 78–99, Madison, WI, USA. Association for Computing Machinery.

Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. 2024. Consent in crisis: The rapid decline of the AI data commons. In *Advances in Neural Information Processing Systems 37*, volume 37, pages 108042–108087, Vancouver, Canada. Neural Information Processing Systems Foundation.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems 37*, volume 37, pages 30811–30849, Vancouver, Canada. Neural Information Processing Systems Foundation.

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 *Proceedings of the Second Conference on Language Modeling (COLM 2025)*, Montreal, Canada.

Nicolas Steinacker-Olsztyn, Devashish Gosain, and Ha Dao. 2026. Is misinformation more open? A study of robots.txt gatekeeping on the web. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*, pages 7507–7516, Dubai, UAE. Association for Computing Machinery.

W3C Text and Data Mining Reservation Protocol Community Group. 2024. TDM reservation protocol (TDMRep). Community group final report, W3C.

Appendix A. Per-language tables

Language	Tier	census	N retained	dropped	w_{tail}	head % of mass	doc mass (unwt.)	doc mass (design-wt.)
English	high	137,447	20,000	117,447	113.3	10.9	62.4	39.2
German	high	54,729	20,000	34,729	113.3	36.0	52.7	42.6
Japanese	high	102,783	20,000	82,783	113.3	33.2	23.6	14.3
French	high	59,428	20,000	39,428	113.3	31.5	38.6	27.1
Indonesian	mid	92,687	20,000	72,687	113.3	18.3	16.0	8.4
Turkish	mid	51,349	20,000	31,349	113.3	38.2	14.6	8.8
Vietnamese	mid	67,868	20,000	47,868	113.3	28.3	8.3	6.2
Thai	mid	68,925	20,000	48,925	113.3	24.7	17.3	16.3
Azerbaijani	mid	14,838	14,838	0	83.8	57.3	10.5	9.5
Swahili	low	9,225	9,225	0	51.7	68.6	11.1	9.4
Sundanese	low	10,339	10,339	0	58.1	55.2	8.2	7.7
Yoruba	low	5,641	5,641	0	31.2	69.3	17.4	15.4
Uyghur	low	3,453	3,453	0	18.7	85.9	10.8	12.3
Irish	low	14,487	14,487	0	81.8	54.0	13.5	12.3
Tier means, document mass							44.3 / 13.3 / 12.2	30.8 / 9.8 / 11.4
Tier means, token mass							38.8 / 11.7 / 11.1	27.6 / 8.7 / 11.2

Table 2: Sample design and composition. Fourteen language subsets contribute 350 domains each across high/mid/low tiers. "census" is the pre-cap domain count, N the retained count, $w_{\text{tail}} = (N - 175)/175$ the tail weight, and "head % of mass" the share in 175 certainty-selected domains; final columns are unweighted and Horvitz–Thompson design-weighted document mass behind a block.

Language	Tier	any-15	CCBot only	GPTBot only	CCBot or $\star$	any-15 or $\star$	$\star$ only
English	high	46.7	40.8	38.2	44.1	49.6	9.6
German	high	29.9	25.2	24.0	34.3	39.0	13.4
Japanese	high	12.8	8.8	10.6	11.5	15.0	4.0
French	high	23.0	20.4	17.8	23.0	25.3	6.7
Indonesian	mid	13.8	12.0	12.4	13.3	14.7	8.9
Turkish	mid	10.9	8.8	10.5	11.3	13.0	8.8
Vietnamese	mid	8.0	7.0	7.5	8.9	9.9	8.0
Thai	mid	12.6	10.7	10.7	13.6	15.4	11.7
Azerbaijani	mid	10.4	9.5	9.9	11.7	12.6	11.3
Swahili	low	5.4	4.9	5.4	7.4	7.9	5.4
Sundanese	low	12.3	10.4	9.2	19.0	20.9	17.2
Yoruba	low	11.3	9.9	7.9	13.9	15.2	10.6
Uyghur	low	10.9	9.4	10.1	18.1	18.8	16.7
Irish	low	10.4	7.9	9.4	10.9	13.4	7.4
Tier means		28.1 / 11.1 / 10.0	23.8 / 9.6 / 8.5	22.7 / 10.2 / 8.4	28.2 / 11.8 / 13.9	32.2 / 13.1 / 15.2	8.4 / 9.7 / 11.5
High vs low p		0.008	0.032	0.008	0.064	0.040	0.37
High vs rest p		0.002	0.007	0.003	0.008	0.005	0.36

Table 3: Domain-level opt-out (%) over status-200 responses under six crawler policies. CCBot only is the Common Crawl rebuild policy; $\star$ is blanket `User-agent: *` with `Disallow: /`. Tier means are high/mid/low, and p-values are exact equal-weighted language-partition permutations.

Language	Tier	reach.	conv. %	AI %	conv. n	AI | conv. %	non-def. n	opt-out | non-def. %
English	high	272	43.8	49.3	119	72.3	242	49.2
German	high	254	39.8	40.2	101	77.2	212	35.4
Japanese	high	226	26.5	15.9	60	33.3	171	15.8
French	high	269	31.2	26.4	84	58.3	217	26.3
Indonesian	mid	225	26.7	27.6	60	68.3	171	17.5
Turkish	mid	238	22.7	21.4	54	53.7	193	13.5
Vietnamese	mid	213	12.2	8.9	26	15.4	179	9.5
Thai	mid	214	13.6	16.8	29	31.0	160	16.9
Azerbaijani	mid	222	18.9	13.1	42	26.2	161	13.7
Swahili	low	203	11.3	6.4	23	30.4	114	9.6
Sundanese	low	163	32.5	12.9	53	13.2	96	16.7
Yoruba	low	151	41.7	11.3	63	14.3	101	13.9
Uyghur	low	138	21.7	10.9	30	16.7	109	13.8
Irish	low	202	20.8	12.4	42	28.6	147	12.2
Pooled high		1,021	35.7	33.6	364	64.0	842	33.0
Pooled mid		1,112	19.0	17.7	211	44.5	864	14.1
Pooled low		857	24.6	10.6	211	19.0	567	13.1
High vs low p			0.190	0.008		0.008		0.016

Table 4: Mechanism proxies. "conv." and "AI" are naming rates for fifteen conventional and AI crawlers; "non-default" excludes stock CMS and trivial bodies. p-values are exact equal-weighted language-partition permutations. Site-type values are pooled high/mid/low: 15.2/10.5/2.3% for `.org`/`.edu`/`.gov` and 31.9/12.4/10.9% for hostnames with no keyword; keyword classes are non-exclusive and some cells have $n < 20$.

Language	Tier	sampled	reachable	reach %	DNS	timeout	403	404/410	5xx	other
English	high	350	272	77.7	14	14	26	11	4	9
German	high	350	254	72.6	24	11	12	37	1	11
Japanese	high	350	226	64.6	57	6	13	37	2	9
French	high	350	269	76.9	30	11	14	17	0	9
Indonesian	mid	350	225	64.3	66	9	16	21	4	9
Turkish	mid	350	238	68.0	56	13	14	17	4	8
Vietnamese	mid	350	213	60.9	76	14	18	18	2	9
Thai	mid	350	214	61.1	78	12	16	19	2	9
Azerbaijani	mid	350	222	63.4	72	5	17	18	3	13
Swahili	low	350	203	58.0	81	11	11	32	2	10
Sundanese	low	350	163	46.6	107	11	23	32	5	9
Yoruba	low	350	151	43.1	107	15	20	41	5	11
Uyghur	low	350	138	39.4	101	15	12	64	3	17
Irish	low	350	202	57.7	64	13	12	44	2	13
Pooled high		1,400	1,021	72.9	8.9%	3.0%	4.6%	7.3%	0.5%	
Pooled mid		1,750	1,112	63.5	19.9%	3.0%	4.6%	5.3%	0.9%	
Pooled low		1,750	857	49.0	26.3%	3.7%	4.5%	12.2%	1.0%	

Table 5: Fetch outcomes for 350 sampled domains per language. "sampled" is total and "reachable" is status-200; remaining columns count failure types, while pooled rows are tier shares of all fetches. "other" combines connection resets, SSL errors, remaining 4xx, and uncommon statuses.

Author Index

Baranwal, Pushkar, 131
Birkenes, Magnus Breder, 139
Burlova, Albina, 282

Cai, Yuedan, 88
Charalampous, Antonis, 8
Cheng, Austin, 347
Chernikov, Oleg, 282
Coats, Steven, 192

Dahal, Ahrnish P, 131
Dinić, Vuk, 53
Djouvas, Constantinos, 8
Dong, Anastasia Zhi-Ling, 200

Flesch, Marie, 182
Foppiano, Luca, 107

Galica, Tomasz, 60, 115, 200
Ginter, Filip, 252
Ginzburg, Jonathan, 331
Goyal, Sharvin, 214, 220
Guan, Enci, 88
Guo, Jiaqi Feng, 200

Henriksson, Erik, 60, 147, 200, 252

Kanerva, Jenna, 252
Kanner, Antti, 147, 200
Karaosmanoğlu, Toprak Seda, 297
Kunnas, Urho, 115, 200

Laippala, Veronika, 60, 115, 147, 200
Li, Yibo, 88
Libovický, Jindřich, 1
Liu, Geng, 88
Ljubešić, Nikola, 53
Lundberg, Tuomas, 147, 200

McCloskey, Vincent, 224

Newman, Christian, 37

Ostendorff, Malte, 70

Pierri, Francesco, 88
Psaltaki, Erofili, 200
Ptitsyn, Evgenii, 282
Pundir, Amit, 131
Pyysalo, Sampo, 115

Raihan, Nishat, 37
Ranasinghe, Tharindu, 308
Razzaghi, Alireza, 60, 200
Rigal, Jacob, 331
Rupnik, Peter, 53

Sarveswaran, Kengatharaiyer, 308
Saxena, Geetika Jain, 131
Singh, Sanjeev, 131
Sivakulasingam, Ajintha, 308
Sobhy, Amr, 22
Srinivasan, Krishna, 224

Tarkka, Otto, 252
Tichý, Michal, 1

Wang, Jessica, 320
Wong, Sidney, 238

Zampieri, Marcos, 37
Zhang, Gary, 161, 170, 347
Zhang, Ruiyang, 100
Zhou, Calvin, 224