FineWeb-CLaR: Culture, Language, and Region Annotations for Benchmark-Aligned Corpus Auditing

Anonymous ACL submission

Abstract

Cultural evaluation coverage and robustness in language models are difficult to diagnose because pretraining corpora and cultural benchmarks are rarely indexed with comparable metadata. Benchmarks increasingly target culturally situated phenomena at the level of languages, regions, and locale-specific practices, while web-scale corpora are usually organized only by language. A shared culture-language-region layer makes these resources comparable, enabling audits of whether a target cultural phenomenon is represented in pretraining data, evaluated by benchmarks or both. We introduce FineWeb-CLaR, a large-scale annotated dataset derived from FineWeb and FineWeb-2 that places web documents on a shared culture–language–region axis for corpus auditing and benchmark alignment.

FineWeb-CLaR annotates the full 5.54B-document collection from FineWeb and FineWeb-2 with URL-derived region labels and cultural-topic provenance. The region resolver assigns a non-empty region to 57.20% of documents, of which 31.76% (1.65B) are retained under conservative lookup-backed criteria, spanning 31,821 language-region locales. For cultural-topic analysis, we focus on the 6,064 locales with at least 500 documents: we induce locale-specific topics and project them onto the 14 leaves of the Cultural Taxonomy of Liu et al. (2025), producing Locale Topic Distributions (LTDs) for corpus-side comparison. We also annotate 277 cultural NLP benchmarks with the same taxonomy, language coverage, and region coverage. Together, these resources enable direct comparison between corpus-side pretraining evidence and benchmark-side evaluation coverage.[1]

[1]Anonymous code and data sample for review: https://anonymous.4open.science/r/FineWeb-CLaR-F018. Full release on Hugging Face on acceptance.

1 Introduction

Cultural evaluation coverage and robustness in language models are difficult to diagnose because failures may stem from missing cultural evidence in pretraining data, weak benchmark coverage, poor alignment with target communities, or model behaviour that relies on surface associations rather than culturally grounded understanding (Adilazuarda et al., 2024; Chiu et al., 2024a; Naous and Xu, 2025; Vo and Koyejo, 2025). Diagnosing these cases requires comparing the cultural phenomena targeted by benchmarks with the evidence available in pretraining data.

Existing web-scale pretraining corpora are not designed for this comparison. Corpora such as FineWeb (Penedo et al., 2024) and FineWeb-2 (Penedo et al., 2025) are organized primarily by detected language, but language is too coarse for culturally situated analysis. Many languages span multiple regions, institutions, communities, and communicative norms. Following the Cultural Taxonomy (CT) of Liu et al. (2025), we treat culture as *socially shared phenomena* organized into ideational, linguistic, and social dimensions, covering knowledge, values, beliefs, norms, practices, language varieties, discourse conventions, social groups, identities, relations, roles, and institutions. Existing language-only corpus annotations cannot show which of these phenomena appear in which regional or locale-specific slices of pretraining data.

To that end, we introduce FineWeb-CLaR, a large-scale annotated dataset derived from FineWeb and FineWeb-2 that places web documents on a shared language-region-cultural-topic axis. The language axis comes from the original corpus; for the region axis we assign URL-derived ISO 3166-1 alpha-2 labels with confidence metadata; and for the cultural-labels axis we map the corpus to the 14 leaves of the CT proposed by Liu et al. (2025).

FineWeb-CLaR is released at two scales. First, we provide URL-derived region annotations for the full 5.54B-document collection. The resolver assigns a non-empty region label to 57.20% of documents, yielding 31,821 language-region locales, of which 31.76% (1.65B documents) carry conservative lookup-backed labels. Second, for the 6,064 locales with at least 500 documents, we induce locale-specific topics with FASTopic (Wu et al., 2024) and project them onto CT leaves to produce Locale Topic Distributions (LTDs). We also annotate 277 cultural NLP benchmarks with the same taxonomy, language, and region coverage, extending prior taxonomy-driven surveys with 212 newly identified benchmarks. Together, these annotations support direct comparison between corpus-side pretraining evidence and benchmark-side evaluation coverage.

Our contributions are:

- We introduce FineWeb-CLaR, a large-scale version of FineWeb and FineWeb-2 augmented with region metadata and CT-aligned topic metadata for culture-, language-, and region-aware corpus auditing.
- We propose a scalable URL-based region attribution method for web-scale data with confidence and resolution-source metadata.
- We derive cultural-taxonomy-aligned locale-level data using topic modeling, embedding-based ranking and targeted LLM refinement.
- We extend prior taxonomy-driven cultural benchmark surveys (Liu et al., 2025; Namazifard and Galke, 2025) by annotating 212 additional cultural NLP benchmarks with locale metadata.

2 Related Work

Culture, benchmarks, and corpus audits. Recent work has argued that NLP systems and evaluations often conflate language, nationality, ethnicity, geography, and culture, while cultural evaluation resources vary in how they operationalise knowledge, values, norms, social relations, and situated behaviour (Hershcovich et al., 2022; Adilazuarda et al., 2024; Liu et al., 2025). We build on the Cultural Taxonomy of Liu et al. (2025), which organizes cultural phenomena into ideational, linguistic, and social dimensions, and use it as a shared coding scheme for corpus-side topic metadata and benchmark-side target phenomena. Prior benchmark efforts evaluate multicultural factual knowledge, everyday cultural knowledge, values, social bias, stereotypes, and culturally grounded reasoning (Myung et al., 2024a; Chiu et al., 2024a; Durmus et al., 2023; Pistilli et al., 2024; Nangia et al., 2020b; Parrish et al., 2022a). Taxonomy-driven resource surveys and curations, including Liu et al. (2025) and MUREL (Namazifard and Galke, 2025), make these resources more comparable. Our benchmark audit follows this line of work, but connects benchmark coverage to corpus-side evidence by annotating both with the same language–region–taxonomy metadata.

Web corpora and geographic attribution. Large web corpora such as CCNet, OSCAR, mC4, CC100, CulturaX, FineWeb, and FineWeb-2 improve pretraining-data transparency through language identification, filtering, deduplication, and quality-control pipelines, but they are primarily organized by detected language rather than by region or locale (Wenzek et al., 2020; Ortiz Suárez et al., 2020; Xue et al., 2021; Conneau et al., 2020; Nguyen et al., 2023a; Penedo et al., 2024, 2025). Geographically organized web corpora are a step closer to our culture–language–region annotation objectives: the Corpus of Global Language Use indexes Common Crawl data by language-country pairs (Dunn, 2020), and GloWbE organizes web English into country-specific varieties (Davies and Fuchs, 2015). We follow this tradition of URL-based geographic attribution, using ccTLDs, public-suffix-aware domain parsing, host and path tokens, and locale parameters. Because such signals are noisy and do not directly identify author location, audience, or cultural membership, FineWeb-CLaR releases region labels together with confidence and resolution-source metadata.

Topic modeling and cultural mapping. Topic modeling is widely used to summarize large corpora into interpretable thematic distributions, from classical models such as LDA (Blei et al., 2003) to neural methods such as BERTopic and FASTopic (Grootendorst, 2022; Wu et al., 2024). Multilingual topic models have also been used to compare culturally or linguistically distinct corpora (Gutiérrez et al., 2016). Our use of topic modeling differs from work that treats discovered topics as the final analysis object, and from supervised label-aligned topic models that use task labels during training (Yang et al., 2025). Instead, we induce topics inde-

pendently within each language-region locale using multilingual document BGE-M3 embeddings (Chen et al., 2024), and FASTopic (Wu et al., 2024), then project topic centroids post hoc onto the 14 leaves of the Cultural Taxonomy to obtain locale topic distributions.

3 FineWeb-CLaR Corpus Annotation

We describe our two-stage annotation approach for labeling FineWeb (Penedo et al., 2024) and FineWeb-2 (Penedo et al., 2025) with region information and mapping the resulting language-region locales to the Cultural Taxonomy of Liu et al. (2025).

3.1 Region Attribution

Cultural benchmarks often specify evaluation targets at the level of countries, regions, dialect areas, or language–region communities which is missing from pretraining data that are organized primarily by language. We close this gap and add document-level region metadata to FineWeb and FineWeb-2 so that pretraining evidence can be aggregated into units closer to those used by cultural evaluations.

3.1.1 Region Labels

A region label can refer to several different concepts, including author location, server location, target audience, domain jurisdiction, or the country conventionally associated with a language. We use a narrower operational definition: for a URL u, a region label r means that URL-derived evidence supports an association between u and the ISO 3166-1 alpha-2 region r (International Organization for Standardization, 2026). Prior work has shown that web-scale corpora can be organized geographically using URL-, domain-, and TLD-derived evidence (Dunn, 2020), but ccTLD and public-web signals provide only partial and source-dependent coverage of local web spaces (Sommese et al., 2023). We therefore use URL-derived region labels only as weak auditing metadata, not as author-location, audience-location, or cultural-membership labels.

This conservative definition is necessary because URL evidence is scalable but noisy. URLs and crawl metadata are available for web-scale corpora and are preserved in Common Crawl WARC/WAT records (Common Crawl, 2023). Regional signals such as ccTLDs, hostnames, path tokens, query parameters, and language headers can provide evidence of regional association, but they are imperfect: domains may be used outside their nominal jurisdictions, generic TLDs such as `.com` are widespread, and hosting or domain choices do not directly identify authorship or cultural affiliation. We therefore release region labels together with source and confidence metadata.

3.1.2 Resolver Design

The region resolver separates offline region evidence aggregation from corpus-scale annotation and emits three fields for each document: a region label, a resolution source, and a confidence value. The region label is an ISO 3166-1 alpha-2 code or `XX` when no region can be resolved; the resolution source records which URL signal produced the label; and the confidence value records the resolver's confidence in that assignment.

In the offline stage, we build a frozen URL-signature lookup table from Common Crawl records sampled across the available `CC-MAIN` archives, ranging from `CC-MAIN-2008-2009` to `CC-MAIN-2026-08`. Each sampled record contributes a URL and, when available, HTTP `Content-Language`, HTML `lang`, and language-identification metadata from FineWeb. We extract URL signatures at three granularities: registered domain, full host, and host plus first path segment, then aggregate weighted evidence over candidate regions for each signature. Each lookup row stores the winning region, confidence tier, and supporting metadata; details of the evidence rules are provided in Appendix B.

In the search-time stage, we apply the frozen lookup table to every FineWeb and FineWeb-2 document. Given a URL, the resolver follows the fixed priority chain shown in Table 1; the first source that emits a valid label determines the document's region and resolution source. Lookup-backed matches inherit the confidence metadata stored in the frozen lookup table, while fallback sources receive separate weak confidence labels. Details of the search-time resolution sources are provided in Appendix C. The production pass never inspects document text, which keeps region attribution independent of the cultural-topic analysis in Section 3.2.

3.1.3 Confidence and Coverage Settings

Our lookup table contains 1,335,000 URL-signature rows. Confidence is assigned after evidence aggregation rather than per individual crawl record, and falls into three tiers (`high`, `medium`, `low`) combining vote count, vote share, and dump

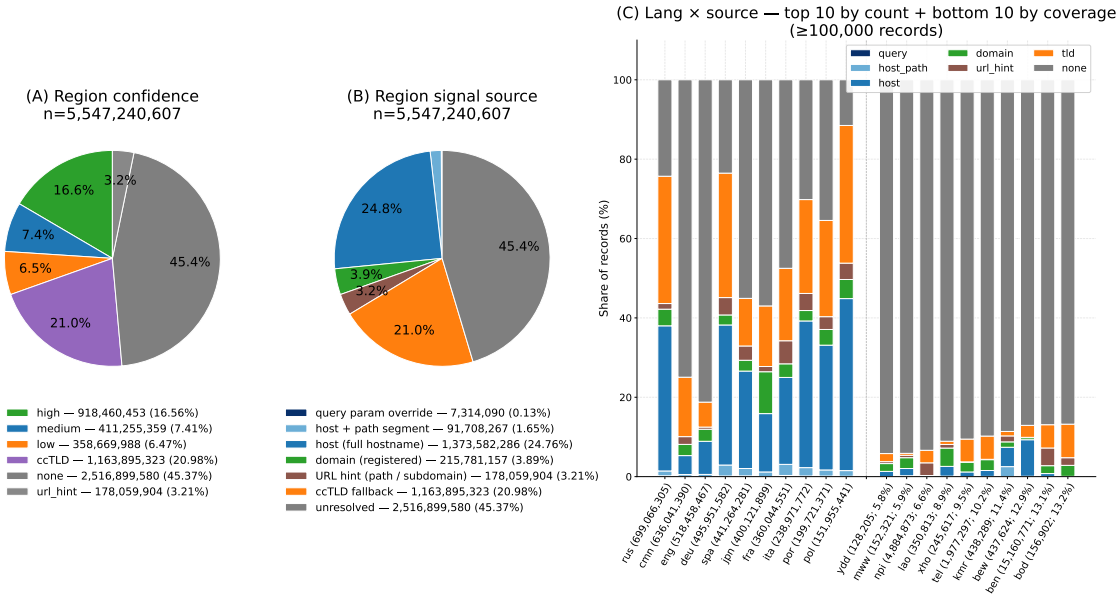

Figure 1: Region-resolution diagnostics: (A) confidence values, (B) resolution sources, and (C) source composition for the largest and lowest-coverage languages.

#	Source	Signal	Conf
1	`query`	explicit locale parameter	`high`
2	`host_path`	lookup table	inherited
3	`host`	lookup table	inherited
4	`domain`	lookup table	inherited
5	`url_hint`	curated URL token	`url_hint`
6	`tld`	non-branded ccTLD	`ccTLD`
7	`none`	no region signal	`none`

Table 1: Search-time priority chain for region resolution, with lookup-backed sources inheriting table confidence and fallback sources receiving weak confidence labels.

diversity; exact thresholds are given in Appendix B.

We report two coverage settings. The *broad* setting retains any non-`XX` region emitted by the resolver, including weak URL-hint and ccTLD fallback labels. The *conservative* setting retains only lookup-backed labels with inherited `high`, `medium`, or `low` confidence. Thus, `host_path`, `host`, and `domain` matches are eligible for the conservative setting, while `url_hint` and `tld` increase recall but are treated as weak evidence. This separation allows downstream users to trade off recall and precision according to the requirements of their analysis. The empirical distribution of these resolution sources is reported in Figure 1 as a diagnostic for interpreting broad and conservative coverage.

3.1.4 Resolver Calibration Against Independent References

Defining a single gold-standard region label for an arbitrary URL is difficult: the legal jurisdiction of the registrant, the physical hosting location, the content-declared locale, and the intended audience may all disagree (e.g., Egyptian-recipe content written in Arabic, hosted by a French CDN on a generic `.com` domain, with `og:locale=ar_EG`). We therefore do not have a single reference standard for evaluation, and human annotations are impractical at this scale. Instead, we report *convergence* between the resolver's output and two independent automatic reference signals, each capturing a different notion of "region":

- **IP geolocation** (MaxMind GeoLite2 (MaxMind, Inc., 2026)), capturing physical hosting jurisdiction. For each URL we resolve the hostname to an IP address and look up its country in GeoLite2, yielding the jurisdiction of the server that serves the bytes.
- **Content-declared locale**, captured from the HTML `lang` attribute (Phillips and Davis, 2009) and the Open Graph `og:locale` property (Facebook, 2014). For each URL we re-fetch the page, parse the root `<html lang=". . . ">` attribute and any `og:locale` meta tag, and take the region subtag (e.g. `en-GB` or `ar_EG` → region `GB` or `EG`).

For each confidence tier of the resolver, Table 2 reports the percentage of URLs whose label agrees with each reference, the percentage where the resolver and both references coincide, and the percentage where all three sources disagree — which

Tier	N	IP	og:locale	All agree	All disagree
high	1000	21.9	41.4	9.1	14.4
medium	1000	19.8	16.1	3.6	19.0
low	1000	17.1	9.6	3.8	22.6
broad-only	1000	37.8	17.1	9.3	33.2

Table 2: Resolver calibration against independent automatic references on stratified random samples of 1,000 URLs per tier. Cell values are percentages of URLs in the tier. The table reports convergence between three distinct notions of "region": the resolver's audit-unit label, IP-based hosting jurisdiction, and content-declared `og:locale`.

CT branch	CT leaves used in this work	Sub-leaves
Ideational	Concepts	—
	Knowledge	—
	Norms and Morals	—
	Artifacts	—
	Values	General; Bias; Hate; Other Perceptions
Linguistic	Dialects	—
	Styles, Registers and Genres	—
Social	Relationship	—
	Context	—
	Communicative Goals	—
	Demographics	—

Table 3: Cultural Taxonomy label space used for corpus-side topic alignment, adapted from Liu et al. (2025). We use 14 leaf categories ($C = 14$) across three branches.

we interpret as URLs that are genuinely ambiguous rather than as resolver failures.

Two patterns are worth highlighting. First, the all-three-disagree fraction increases monotonically with decreasing resolver confidence (14.4%, 19.0%, 22.6% across `high`, `medium`, `low`; rising to 33.2% for `broad-only` URLs that have no lookup-backed evidence at all), indicating that the confidence tiers correctly identify which URLs are inherently ambiguous across reference signals. Second, agreement with `og:locale` drops sharply across tiers (41.4% $\rightarrow$ 16.1% $\rightarrow$ 9.6%), confirming that high-tier URLs are also the ones whose content most frequently declares the same locale. Agreement with IP geolocation is broadly tier-invariant ($\sim$17–22%) because IP measures physical hosting, which is largely orthogonal to the resolver's audit-unit semantics; the higher 37.8% for `broad-only` reflects ccTLD- and hint-based URLs whose region label trivially aligns with hosting jurisdiction. Both reference signals are themselves imperfect: IP geolocation captures hosting rather than authorship, and `og:locale` is content-declared rather than verified. We therefore treat this analysis as a convergence study rather than a validation against ground truth, and recommend that downstream users rely on the released confidence tiers when interpreting individual region labels.

3.2 Culture-Taxonomy Annotation

The region axis introduced in Section 3.1 identifies where a document plausibly originates from, but it does not identify what kind of cultural phenomenon is present in the document or its local corpus context. We therefore add cultural-topic labels to the region-resolved corpus.

3.2.1 Taxonomy and Label Space

We operationalize the Cultural Taxonomy (CT) of Liu et al. (2025) as the cultural label space. The CT organizes cultural phenomena into ideational, linguistic, and social dimensions. Table 3 summarizes the label space used for topic alignment. The Values element of the Ideational branch is further refined into four sub-leaves, yielding $C = 14$ leaf categories in total. The goal of our mapping is to estimate which CT categories are evidenced by a document's topic context.

3.2.2 Locale-Specific Topic Induction

Human document-level CT annotation at the scale of FineWeb and FineWeb-2 is infeasible. We therefore use a topic pipeline over language–region locales. Although the full region-resolved FineWeb-CLaR inventory contains 31,821 language–region locales, not all locales contain enough documents for stable topic modeling; we retain locales with at least 500 documents after filtering. This yields 6,064 topic-modeled locales spanning 307 languages (336 distinct language–script combinations). The resulting culturally annotated resource is therefore a subset of the full region-annotated corpus.

For each retained locale, we encode documents using multilingual BGE-M3 (Chen et al., 2024) and induce topics independently with FASTopic (Wu et al., 2024). Modeling each locale separately prevents high-resource languages and globally dominant regions from determining the topic space for smaller locales, and allows locale-specific cultural content to form separate clusters.

We use a fixed initial maximum number of topics $K = 200$ for every retained locale. This keeps the topic-induction hyperparameter constant across locales, which is important for the cross-locale comparisons in Section 5.3.[2]

[2]The fixed value should be interpreted as an overcomplete topic budget rather than as the final number of topics: the effective topic count is determined after the merge step.

FASTopic (Wu et al., 2024) returns a document-topic distribution and topic descriptors. For each topic, we compute a centroid as the mean BGE-M3 embedding of the documents whose highest-probability topic assignment is that topic. We use these centroids for CT projection to the BGE-M3 CT leaf-description embeddings.

The fixed topic budget can produce near-duplicate topics, especially in smaller locales. We therefore merge topics using agglomerative average-linkage clustering from scikit-learn (Pedregosa et al., 2011) over cosine distance between topic centroids, cutting clusters at cosine similarity ≥ 0.85. For each merged topic, we sum the corresponding columns of the document-topic matrix, compute a document-count-weighted centroid, and retain the top unique words from the merged member topics. The merged topics are then passed to the topic-to-taxonomy projection step.

3.2.3 Topic-to-Taxonomy Projection

We project topic centroids onto CT-leaf embeddings (both encoded with BGE-M3) via a temperature-scaled cosine softmax with $\tau = 0.05$; full equation and rationale in Appendix D.

3.2.4 Low-Confidence Adjudication

The embedding-based projection is used as the default topic-to-CT mapping. For low-confidence topics, defined as topics whose maximum CT probability is below 0.5, we apply targeted LLM adjudication. The adjudicator receives the topic descriptors together with the top-three CT candidate categories and their softmax probabilities, and either confirms the automatic label or selects a better-fitting CT leaf. Table 4 shows example FASTopic topics and their final CT leaf assignments for selected locales.

Locale	Topic	CT leaf
en-US	Christian Religion	Values-General
en-US	Restaurant Dining	Concepts
en-US	US Politics	Demographics
ar-EG	Travel Destinations	Concepts
ar-EG	Sports Media	Context
zh-CN	Digital Media & Devices	Artifacts
hi-IN	Indian Culture	Knowledge
ko-KR	Korean Global Education	Demographics

Table 4: Example FASTopic topics and final CT leaf assignments for selected locales.

3.2.5 Locale Topic Distributions

The output of the corpus-side cultural-topic pipeline is a Locale Topic Distributions (LTD) for each topic-modeled locale. For a locale with N documents, K merged topics, and C CT categories, let $D \in R^{N \times K}$ be the document–topic distribution and $A \in R^{K \times C}$ be the topic-to-CT assignment matrix. The locale's Locale Topic Distribution is:

$$\text{LTD} = \frac{1}{N} \sum_{i=1}^{N} (DA)_i.$$

The resulting vector is L1-normalized and interpreted as the locale's distribution over CT leaves. Across all topic-modeled locales, this procedure produces a $6{,}064 \times C$ matrix of LTD, with $C = 14$. We use this matrix for distributional analyses of locale-level cultural-topic profiles, including the same-language regional divergence analysis in Section 5.3.

4 Benchmark Audit

To compare corpus-side cultural evidence with evaluation demand, we construct a benchmark-side audit using the same Cultural Taxonomy label space and locale metadata as FineWeb-CLaR. The audit extends prior taxonomy-driven cultural-NLP surveys (Liu et al., 2025; Namazifard and Galke, 2025) by adding newly identified datasets and by coding each resource for cultural, language, and region coverage, provenance, and usability.

Dataset collection. We compile cultural NLP benchmarks from three sources: datasets included in the cultural-NLP taxonomy and survey of Liu et al. (2025), datasets included in MUREL-style cultural-resource curation (Namazifard and Galke, 2025), and newly identified resources from systematic searches of the ACL Anthology, arXiv, HuggingFace, GitHub, and dataset repositories. After deduplication and filtering, the audit contains 277 cultural NLP benchmark datasets (43 inherited from the Liu et al. taxonomy survey, 3 from MUREL, 19 shared by both, and 212 newly identified). A resource is included when it provides an accessible dataset, targets at least one cultural phenomenon covered by the CT, and is documented well enough to recover its language, region, task, and provenance metadata. Figure 2 shows how the expanded audit changes CT-leaf coverage relative to the original taxonomy survey and MUREL-only additions. The newly identified datasets increase coverage across most CT leaves, including linguistic and social categories that are comparatively sparse in the inherited survey sources.

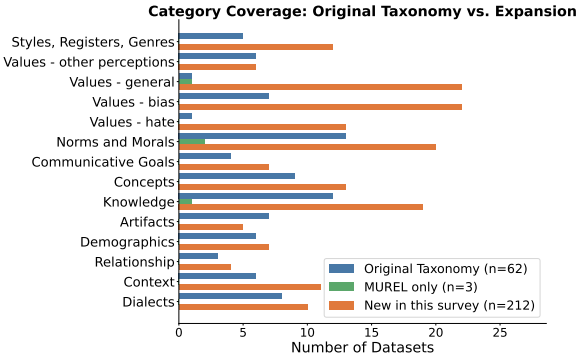

Figure 2: Benchmark coverage by CT leaf and inclusion source, comparing inherited survey datasets and newly added resources across the 14 CT leaves.

Annotation schema. Each benchmark is annotated along four groups of fields. Cultural coverage records the CT branch and one or more CT leaves targeted by the dataset. Scope records the covered languages, regions, task format, and example items when available. Provenance records the source paper, venue, publication date, generation method, source list, and dataset lineage. Usability records whether the dataset is public, restricted, or unavailable.

Human annotation and validation. Two authors annotated the benchmarks and resolved disagreements through discussion. CT branch and CT leaf labels were assigned using each dataset's paper, dataset card, repository, and example items. A dataset may receive multiple CT leaves when it evaluates more than one cultural phenomenon. Details of the search protocol, annotation rubric, lineage coding, accessibility coding, and quality-control procedure are provided in Appendix E.[3]

5 Experiments and Analysis

We evaluate FineWeb-CLaR as a corpus-auditing resource through three analyses: human annotation, corpus-benchmark mismatch, and same-language regional divergence.

5.1 Validation of the Topic-to-CT Projection

Two authors independently coded 100 (topic, CT-leaf) pairs sampled stratified across the 14 CT leaves and across high-, medium-, and low-resource locales. Each pair was presented with the top-15 FASTopic descriptors, ten randomly drawn documents from the topic, and the natural-language definition of each CT leaf. Annotators selected the single best-fitting CT leaf or `NONE`. Annotators did not see the embedding-softmax or LLM-adjudicated labels during annotation.

Cohen's κ between annotators was 0.25, indicating only fair agreement on the topic-to-CT-leaf task and confirming that CT-leaf assignment at the topic level is intrinsically ambiguous. On 15% of topics the two annotators could not converge on a single leaf; the remaining $N = 85$ topics form the consensus reference set. Against this reference, the embedding-softmax label agreed on 18.8% of cases (Cohen's κ = 0.13); the LLM adjudicator, which was triggered on all sampled topics and altered the embedding-softmax assignment for 35% of them, agreed on 22.4% of cases (Cohen's κ = 0.15). Both automatic methods are well below human consensus, and the adjudicator yields only marginal improvements over the embedding baseline. We therefore interpret LTDs not as definitive CT-leaf assignments but as locale-level distributions over topic-to-leaf *prevalences* under substantial labeling uncertainty.

5.2 Corpus–Benchmark Mismatch

The shared language–region–CT axis allows us to compare corpus evidence with benchmark demand. On the corpus side, FineWeb-CLaR provides document counts and Locale Topic Distributions for language–region locales. On the benchmark side, the human annotation in Section 4 provides CT leaves, language coverage, and region coverage for 277 cultural NLP benchmarks.

We mark a CT leaf–region cell as corpus-covered when at least $k = 50$ documents are associated with that cell. Each cell is then classified as both present, corpus only, benchmark only, or neither present, depending on whether it appears in FineWeb-CLaR, the benchmark inventory, both, or neither.

Figure 3 shows that mismatch is structured by both region and CT branch. Sparse coverage concentrates in lower-resource regions and in linguistic categories such as *Dialects* and *Styles, Registers, and Genres*. Ideational categories such as *Knowledge*, *Concepts*, and *Artifacts* are more often visible in web text and more often converted into benchmark tasks. Thus, corpus–benchmark mismatch is not only a language-level problem; it also depends on which cultural phenomenon is being evaluated.

[3]We release the full benchmark survey in our code repository.

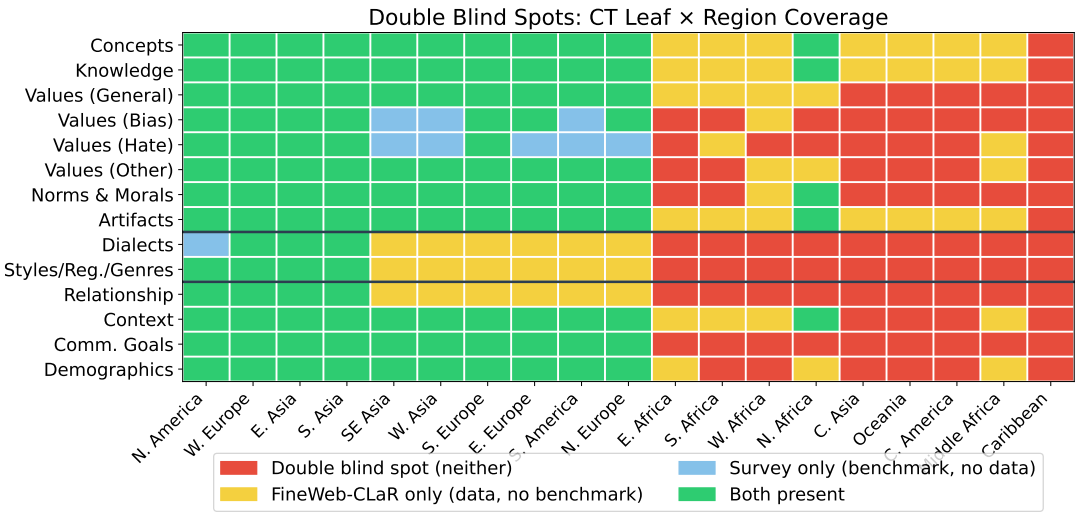

Figure 3: Coverage gaps over CT leaf × region cells. Cells are corpus-covered if at least $k = 50$ documents support the CT-leaf/region pair, and benchmark-covered if at least one audited benchmark targets it.

5.3 Same-Language Regional Divergence

Language-level corpus partitions can hide regional variation. We quantify this by computing the Jensen–Shannon distance (JSD; full definition in Appendix D) between Locale Topic Distributions for locales that share a language but differ in region. Higher JSD indicates greater divergence between the two locale-level CT distributions.

To distinguish genuine regional divergence from topic-modeling noise, we compute a within-locale split-half JSD baseline. For each of the 42 locales with at least 1,000 documents, we randomly partition the locale's documents into two halves, fit an LTD on each half, and compute the JSD between halves; we repeat this 50 times per locale (2,100 split-half computations in total, seed 42). The pooled distribution yields a mean JSD of 4.2×10^{-4} (SD = 1.0×10^{-3}, $p_{95} = 1.4 \times 10^{-3}$), which we use as the topic-model noise floor.

Figure 4 shows the distribution of within-language JSD across every language with at least two locales meeting the 500-document floor. The labelled examples all lie well above the split-half noise floor, indicating that same-language regional differences in LTDs are substantially larger than topic-modeling noise. The largest labelled divergences occur for Arabic and Chinese locale pairs, followed by English, while the Portuguese and Spanish pairs show smaller but still clearly non-noise divergence. We interpret this as evidence that benchmarks targeting a language may correspond to specific regional CT profiles, while noting that the magnitude of the language-level/regional gap varies substantially across language families.

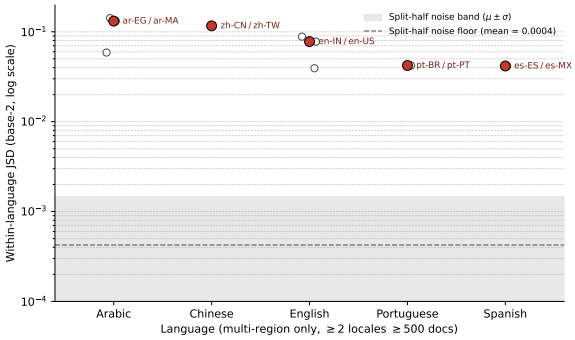

Figure 4: Within-language LTD divergence (JSD) across all languages with ≥ 2 locales. Labelled points show illustrative same-language locale pairs. The dashed line shows the split-half noise floor.

6 Conclusion

We introduced FineWeb-CLaR, a large-scale resource that augments FineWeb and FineWeb-2 with region metadata and Cultural Taxonomy-aligned topic annotations to support culture-, language-, and region-aware corpus auditing. By combining URL-derived region attribution, locale-specific topic modeling, and an expanded audit of cultural NLP benchmarks, FineWeb-CLaR enables direct comparison between pretraining evidence and evaluation demand. Our analyses show that cultural coverage gaps are structured not only by language, but also by region and cultural phenomenon, highlighting the need for benchmark development and corpus auditing methods that move beyond language-only metadata.

Limitations

FineWeb-CLaR is an auditing resource, not demographic or cultural ground truth. Its language, region, and CT axes make corpora and benchmarks comparable, but they do not define cultures or communities. Region labels are derived from URL evidence and should not be interpreted as author location, user location, national identity, or cultural membership. They may be affected by multi-region websites, generic or vanity domains, CDNs, URL parsing errors, branded ccTLDs, and the use of a sampled Common Crawl lookup table. False positives and false negatives are therefore unavoidable.

The CT annotation pipeline is also coarse. CT labels support corpus-level comparison, but they cannot capture the full complexity of cultural practice, identity, power, tacit knowledge, interactional norms, or intra-community variation. Locale-specific topics may reflect domain composition, crawl artefacts, or institutional web presence rather than cultural salience alone. Embedding-based CT ranking and LLM-assisted adjudication reduce annotation cost, but can introduce errors, especially for low-resource languages, mixed-language documents, and culturally specific concepts that are weakly represented in the embedding model or adjudicator.

Locale Topic Distributions are available only for locales with sufficient document support. Locales below the 500-document threshold remain region-annotated, but do not receive LTDs because their topic distributions would be dominated by sampling noise. Even among the 6,064 topic-modeled locales, smaller locales should be interpreted with greater uncertainty.

The benchmark survey is time-bounded and depends on the availability and documentation of existing datasets. Human CT coding reduces noise, but does not establish that a benchmark validly represents the community or cultural phenomenon it claims to test. The coverage matrix should therefore be read as a diagnostic of available resources, not as a normative statement about which cultures or cultural dimensions matter.

Ethics Statement

This work annotates public web-corpus metadata for auditing and analysis. Because region labels can be misused if treated as demographic labels, we release confidence and source metadata and recommend against using weak labels for claims about people, identity, or community representation. Cultural-taxonomy annotations should likewise be used to identify broad coverage patterns, not to essentialise cultures or reduce communities to fixed attributes. Downstream use for data filtering, benchmark construction, or model evaluation should document the selected confidence threshold, the intended interpretation of region labels, and the limitations of URL-derived evidence.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. *Preprint*, arXiv:2406.18682.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Muhammad Dehan Al Kautsar, Aswin Candra, Muhammad Alif Al Hakim, Maxalmina Satria Kahfi, Fajri Koto, Alham Fikri Aji, and 1 others. 2025. Seadialogues: A multilingual culturally grounded multi-turn dialogue dataset on southeast asian languages. *Preprint*, arXiv:2508.07069.

Lama Ayash, Hassan Alhuzali, Ashwag Alasmari, and Sultan Aloufi. 2025. SaudiCulture: A benchmark for evaluating large language models cultural competence within Saudi Arabia. *Preprint*, arXiv:2503.17485.

Valentin Barriere, Nahuel Gomez, Leo Hemamou, Sofia Callejas, and Brian Ravenet. 2025. StandUp4AI: A new multilingual dataset for humor detection in stand-up comedy videos.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 52–57, Turin, Italy.

Federico Becattini, Pietro Bongini, Luana Bulla, Marco Bertini, Ludovica Marinucci, Misael Mongiovì, and

Aldo Gangemi. 2023. VISCOUNTH: A large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(6):1–20.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural lenses on emotion – benchmarking LLMs for cross-cultural emotion understanding. *Preprint*, arXiv:2503.10688.

Mukul Bhutani, Kathleen C. Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Minh Duc Bui, Katharina von der Wense Fabian David Schmid, Joan Plepi, and 1 others. 2025. Multi3Hate: Multimodal, multilingual, and multicultural hate speech detection with vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025. High-dimension human value representation in large language models. 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 5303–5330, Albuquerque, New Mexico. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2023. Cultural adaptation of recipes. *Preprint*, arXiv:2310.17353.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Calvin Chen and 1 others. 2023. C-MFD2.0: Developing a Chinese moral foundation dictionary. *Computational Communication Research*, 5(2).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024a. CulturalBench: A robust, diverse, and challenging cultural benchmark by human-ai cultural teaming. *Preprint*, arXiv:2410.02677.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024b. CulturalBench: A robust, diverse, and challenging cultural benchmark by human-AI CulturalTeaming. *Preprint*, arXiv:2410.02677.

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. *Preprint*, arXiv:2003.05002.

Common Crawl. 2023. Web archiving file formats explained. `https://commoncrawl.org/blog/web-archiving-file-formats-explained`. Accessed 2026-05-22.

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. Introduces the CC100 corpus and XLM-R.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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*, 36(1):1–28.

Nicholas Deas, Elsbeth Turcan, Iván Pérez Mejía, and Kathleen McKeown. 2024. MASIVE: Open-ended affective state identification in English and Spanish. *Preprint*, arXiv:2407.12196.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. See also CLEF 2025 PAN shared task on multilingual text detoxification.

Swapnil Dhanwal, Hritwik Dutta, Hitesh Nankani, Nilay Shrivastava, Yaman Kumar, Junyi Jessy Li, and 1 others. 2020. An annotated dataset of discourse modes in Hindi stories. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1191–1196, Marseille, France. European Language Resources Association.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388. Conference on Language Modeling (COLM) 2024.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Ashutosh Dwivedi, Siddhant Shivdutt Singh, and Ashutosh Modi. 2025. EtiCor++: Towards understanding etiquettical bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9355–9376, Vienna, Austria. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elena V. Epure, Guillaume Salha, Manuel Moussallam, and Romain Hennequin. 2020. Modeling the music genre perception across language-bound cultures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Facebook. 2014. The Open Graph protocol. `https://ogp.me`. Accessed: 2026-05-22.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and 1 others. 2024. DIALECTBENCH: A NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *Preprint*, arXiv:2204.08582.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Rahul Goel, Waleed Ammar, Aditya Gupta, Siddharth Vashishtha, Motoki Sano, Faiz Surani, Max Chang, HyunJeong Choe, David Greene, Kyle He, Rattima Nielsen, Jack Robichaud, Kanika Sethi, Nem Sun, Brian Tan, and Jiexin Zhang. 2023. PRESTO: A multilingual dataset for parsing realistic task-oriented dialogs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *Preprint*, arXiv:2203.05794.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. *Preprint*, arXiv:2504.05154.

E. D. Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. 2021. Hope speech detection in under-resourced kannada language. *Preprint*, arXiv:2108.04616.

Paul J. Hanges and Marcus W. Dickson. 2004. The development and validation of the GLOBE culture and leadership scales. In Robert J. House, Paul J. Hanges, Mansour Javidan, Peter W. Dorfman, and Vipin Gupta, editors, *Leadership, Culture, and Organizations: The GLOBE Study of 62 Societies*, pages 122–151. Sage Publications.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. 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 211–226, Mexico City, Mexico. Association for Computational Linguistics.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023. Comparing styles across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cheng, Antoine Doucet, and 1 others. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

David G. Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind*, 3rd edition. McGraw-Hill, New York, NY.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations Twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Robert J. House, Paul J. Hanges, Mansour Javidan, Peter W. Dorfman, and Vipin Gupta. 2004. *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Sage Publications.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi^{3}WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Tianyi Hu, Maria Maistro, and Daniel Hershcovich. 2024. Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Behavior in the Wild. 2024. Align Via Actions (AVA): Learning behavior aligns LLMs with human opinions in zero-shot. `https://behavior-in-the-wild.github.io/align-via-actions.html`. Project page; full author list pending publication.

Ronald Inglehart and Christian Welzel. 2010. Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics*, 8(2):551–567. Methodology behind the World Values Survey.

International Organization for Standardization. 2026. ISO 3166: Country codes. `https://www.iso.org/iso-3166-country-codes.html`. Accessed 2026-05-22.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2023. KoBBQ: Korean bias benchmark for question answering. *Preprint*, arXiv:2307.16778.

Raviraj Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Long, Sanjay Singh Chauhan, and Niranjan Wartikar. 2025. CultureGuard: Towards culturally-aware dataset and guard model for multilingual safety applications. *Preprint*, arXiv:2508.01710.

Saman Sarker Joy. 2025. BnMMLU: Measuring massive multitask language understanding in Bengali. *Preprint*, arXiv:2505.18951. Findings of ACL 2026.

Michelle Kappl. 2025. Are all Spanish doctors male? evaluating gender bias in German machine translation. *Preprint*, arXiv:2502.19104.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. LLM-GLOBE: A benchmark evaluating the cultural values embedded in LLM output. *Preprint*, arXiv:2411.06032.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Preprint*, arXiv:2404.16019. NeurIPS 2024 Datasets and Benchmarks.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on

IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. ArabicMMLU: Assessing massive multitask language understanding in Arabic. *Preprint*, arXiv:2402.12840.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468, Dublin, Ireland. Association for Computational Linguistics.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2306.04298.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marko Krajňc, Dušan Krevelj, Maja Aigro, Liina Repo, Valentin Skórska, Aleksi Sahala, Yulin Yang, 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)*. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. *Preprint*, arXiv:1904.11783.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2023. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. *Preprint*, arXiv:2308.16705.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. *Preprint*, arXiv:1910.07475.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. CulturePark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024c. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024d. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. *Preprint*, arXiv:2404.10199. Conference on Language Modeling (COLM) 2024.

Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024e. How well do LLMs identify cultural unity in diversity? *Preprint*, arXiv:2408.05102.

Oliver Li, Mallika Subramaniam, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023a. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. 2023b. PaCE: Unified multimodal dialogue pre-training with progressive and compositional experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13402–13416, Toronto, Canada. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on Vietnamese social media texts. *Preprint*, arXiv:2103.11528.

Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2022. Cross-lingual dialogue dataset creation via outline-based generation. *Preprint*, arXiv:2201.13405.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2021. Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European languages. *Preprint*, arXiv:2108.05927.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A benchmark dataset for explainable hate speech detection. *Preprint*, arXiv:2012.10289.

MaxMind, Inc. 2026. GeoLite2 free geolocation data. `https://dev.maxmind.com/geoip/geolite2-free-geolocation-data`. Accessed: 2026-05-22.

Margaret Mitchell, Shaily Bhattacharyya, Sunipa Dev, Yacine Jernite, Sara Hassan, Zeerak Talat, Laura Cabello, and 1 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Youssef Mohamed, Runjia Khan, Khadija Almazrouei, Khadija Khan, Muhammad Abdul-Haq Khattak, Esther Bonnah, David Adelani, Muhammad Khattak, Mohamed Elhoseiny, and Kenneth Church. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023a. Global voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Sourabrata Mukherjee, Akanksha Bansal, Pritha Majumdar, Atul Kr. Ojha, and Ondřej Dušek. 2023b. Low-resource text style transfer for Bangla: Data & models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 34–47, Singapore. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024a. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024b. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics.

Danial Namazifard and Lukas Galke. 2025. Isolating culture neurons in multilingual large language models. *Preprint*, arXiv:2508.02241. Introduces the MUREL cultural dataset.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020a. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhasin, and Samuel R. Bowman. 2020b. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. 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 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. *Preprint*, arXiv:2406.07243.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association (ELRA).

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023a. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023b. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917. ACM.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406. Association for Computational Linguistics.

Takumi Ohashi, Tsubasa Nakagawa, and Hitoshi Iyatomi. 2024. Extended Japanese commonsense morality dataset with masked token and label enhancement. *Preprint*, arXiv:2410.09564.

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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. Introduces the OSCAR corpus.

Ibrahim Tolga Öztürk, Rostislav Nedelchev, Christian Heumann, Esteban Garces Arias, Marius Roger, Bernd Bischl, and Matthias Aßenmacher. 2023. How different is stereotypical bias across languages? *Preprint*, arXiv:2307.07331.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022a. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022b. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozen Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. SemEval-2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 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*, volume 37.

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.

Addison Phillips and Mark Davis. 2009. Tags for identifying languages. RFC 5646.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. CIVICS: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1132–1144.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, and 1 others. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, and 1 others. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376. Association for Computational Linguistics.

Ashmari Pramodya, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. Translating movie subtitles by large language models using movie-meta information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Association for Computational Linguistics.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2022. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. *Preprint*, arXiv:2212.10409.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. NormAd: A framework for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

Manikandan Ravikiran and Subbiah Annamalai. 2021. DOSA: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Luis Rei, Dunja Mladenić, Mareike Dorozynski, Franz Rottensteiner, Thomas Schleider, Raphaël Troncy, Jorge Sebastián Lozano, and Mar Gaitán Salvatella. 2023. Multimodal metadata assignment for cultural heritage artifacts. *Multimedia Systems*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *Preprint*, arXiv:2308.01263.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *Preprint*, arXiv:1804.09301.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri, and Debora Nozza.

2024. Can i introduce my boyfriend to my grandmother? evaluating large language models capabilities on Iranian social norm classification. *Preprint*, arXiv:2406.09123.

Nihar Ranjan Sahoo, Pranamya Kulkarni, Narjis Asad, Arif Ahmad, Tanu Goyal, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics.

Pritish Sahu, Anirudh Som, Dimitra Vergyri, and Ajay Divakaran. 2025. MINDS: A cross-cultural dialogue corpus for social norm classification and adherence detection. In *Findings of the Association for Computational Linguistics: IJCNLP 2025*. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Association for Computational Linguistics.

Shalom H. Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1).

Kirill Semenov and Rico Sennrich. 2025. Measuring the effect of disfluency in multilingual knowledge probing benchmarks. *Preprint*, arXiv:2510.15115.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Castro Riemer, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Antoine Bosselut, Alice Oh Smith, Leshem Choshen Foerster, Rakesh Vassilieva, Sara Hooker, and Marzieh Fadaee. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, and 1 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Raffaele Sommese, Roland van Rijswijk-Deij, and Mattijs Jonker. 2023. This is a local domain: On amassing country-code top-level domains from public data. *Preprint*, arXiv:2309.01441.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. KMMLU: Measuring massive multitask language understanding in Korean. *Preprint*, arXiv:2402.11548.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38. AAAI Press.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Marco Stranisci, Michele De Leonardis, Cristina Bosco, and Viviana Patti. 2021. The expression of moral values in the Twitter debate: a corpus of conversations. *Italian Journal of Computational Linguistics (IJCoL)*, 7(1-2):113–132.

Elza Strazda and Gerasimos Spanakis. 2025. Dutch CrowS-Pairs: Adapting a challenge dataset for measuring social biases in language models for Dutch. *Preprint*, arXiv:2507.16442.

Zhewei Sun and Yang Xu. 2022. Tracing semantic variation in slang. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nir Sweed and Dafna Shahaf. 2021. Catchphrase: Automatic detection of cultural references. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexey Tikhonov, Igor Samenko, and Ivan P. Yamshchikov. 2021. StoryDB: Broad multi-language narrative dataset. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.

Aditya Tomar, Nihar Ranjan Sahoo, and Pushpak Bhattacharyya. 2025. BharatBBQ: A multilingual bias benchmark for question answering in the Indian context. *Preprint*, arXiv:2508.07090.

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K. Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi Malekabadi, and Flor Miriam Plaza-del Arco. 2025. MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association.

Francielle Vargas, Jackson Trager, Diego Alves, Surendrabikram Thapa, Matteo Guida, Berk Atil, Daryna Dementieva, Andrew Smart, and Ameeta Agrawal. 2026. Self-explaining hate speech detection with moral rationales. *Preprint*, arXiv:2601.03481.

Truong Vo and Sanmi Koyejo. 2025. CURE: Cultural understanding and reasoning evaluation – a framework for "thick" culture alignment evaluation in LLMs. *Preprint*, arXiv:2511.12014.

Hongru Wang, Min Li, Zimo Zhou, Gabriel Pui Cheong Fung, and Kam-Fai Wong. 2020. KddRES: A multi-level knowledge-driven dialogue dataset for restaurant towards customized dialogue system. *Preprint*, arXiv:2011.08772.

Josiah Wang, Pranava Madhyastha, Josiel Figueiredo, Chiraag Lala, and Lucia Specia. 2021. MultiSubs: A large-scale multimodal and multilingual dataset. *Preprint*, arXiv:2103.01910.

Tian Wang, Shiliang Liu, Yiwen Wang, Ben Wang, Yi Lin, Junjie Wang, Lin Wang, and Hongwei Yu. 2025. McBE: A multi-task chinese bias evaluation benchmark for large language models. *Preprint*, arXiv:2507.02088.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*. 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.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. 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 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024. FASTopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *Advances in Neural Information Processing Systems*, volume 37.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. *Preprint*, arXiv:1610.08914.

Menglin Xia and Emilio Monti. 2021. Multilingual neural semantic parsing for low-resourced languages. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 185–194. Association for Computational Linguistics.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. Exploring multilingual concepts of human values in large language models: Is value alignment consistent, transferable and controllable across languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 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. Introduces the mC4 corpus.

Haoran Yang, Yingxue Li, Xiuyuan Zhang, and Jimmy Lin. 2025. Supervised neural topic modeling with label alignment. *Transactions of the Association for Computational Linguistics*, 13:249–263.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Tao Liu, and Deyi Xiong. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. 2024. SarcNet: A multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14325–14335, Torino, Italia. ELRA and ICCL.

Pardis Sadat Zahraei and Ehsaneddin Asgari. 2025. I am aligned, but with whom? MENA values benchmark for evaluating cultural alignment and multilingual bias in LLMs. *Preprint*, arXiv:2510.13154.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalind, Jureynolds Jiang, Suraj Sharma, Shilin Qu Linhao Luo Lay-Ki Soon Zhuang Li Gholamreza Haffari Suraj Sharma Yiran Ronald Stevens, and Gholamreza Haffari. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117, Mexico City, Mexico. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, and Gholamreza Haffari. 2023. SocialDial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *Preprint*, arXiv:1804.06876.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022a. Value: Understanding dialect disparity in nlu. *Preprint*, arXiv:2204.03031.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023a. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023b. MultiVALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022b. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics.

Muitze Zulaika and Xabier Saralegi. 2025. BasqBBQ: A QA benchmark for assessing social biases in LLMs for Basque, a low-resource language. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics.

A Additional Related Work

Structured cultural axes. The use of explicit cultural dimensions in NLP is motivated by a longer tradition of cross-cultural psychology and social-scientific measurement. Hofstede's organisational-culture model (Hofstede et al., 2010), Schwartz's theory of basic values (Schwartz, 2012), the GLOBE study of 62 societies (House et al., 2004), and the World Values Survey programme (Inglehart and Welzel, 2010) all seek to organize cultural variation along shared dimensions. We do not adopt these frameworks as direct labels for corpus annotation, but they motivate our use of an explicit taxonomy rather than treating culture as interchangeable with language or country. The Cultural Taxonomy of Liu et al. (2025) is better aligned with NLP resources because it covers ideational, linguistic, and social phenomena, including knowledge, values, norms, artifacts, dialects, registers, relationships, contexts, communicative goals, and demographics.

Cultural evaluation resources. Cultural NLP benchmarks differ substantially in what they test and how they define the relevant population or locale. Some resources focus on everyday cultural knowledge and culturally situated commonsense (Myung et al., 2024a; Chiu et al., 2024a); others target values and opinions (Durmus et al., 2023; Pistilli et al., 2024), social bias and stereotyping (Nangia et al., 2020b; Parrish et al., 2022a), or culturally grounded reasoning and adaptation. Survey and curation efforts such as Liu et al. (2025) and Namazifard and Galke (2025) make progress toward comparability by assigning cultural resources to shared categories. Our audit extends this direction by adding newly identified datasets and by recording language, region, provenance, and usability metadata, enabling comparison with corpus-side coverage.

Multilingual web corpora and URL-derived region signals. Existing multilingual and web-scale corpora expose different forms of metadata, but most are designed around detected language and corpus-quality filtering rather than culture- or region-aware auditing. CCNet, OSCAR, mC4, CC100, CulturaX, FineWeb, and FineWeb-2 provide large-scale text collections with language identification and filtering pipelines (Wenzek et al., 2020; Ortiz Suárez et al., 2020; Xue et al., 2021; Conneau et al., 2020; Nguyen et al., 2023a; Penedo et al., 2024, 2025). Geographic web corpora such as the Corpus of Global Language Use (Dunn, 2020) and GloWbE (Davies and Fuchs, 2015) demonstrate that web data can be organized by language–country or regional variety. At the same time, ccTLD and URL-based evidence only partially captures local web spaces and can be affected by generic domains, hosting infrastructure, content-delivery networks, multilingual sites, and branding choices (Sommese et al., 2023). For this reason, FineWeb-CLaR treats region labels as weak audit metadata and releases confidence tiers and resolution sources rather than presenting URL-derived labels as demographic or cultural ground truth.

Topic models for corpus comparison. Classical topic models such as LDA (Blei et al., 2003) established topic modeling as a way to summarize large corpora, and multilingual topic modeling has been used to compare themes across linguistic and cultural communities (Gutiérrez et al., 2016). Neural topic models use pretrained representations to improve scalability and stability in heterogeneous corpora. BERTopic clusters transformer embeddings before deriving topic descriptors (Grootendorst, 2022), while FASTopic provides fast and stable topic induction suitable for large-scale use (Wu et al., 2024). In FineWeb-CLaR, multilingual BGE-M3 embeddings provide a shared representation space across languages (Chen et al., 2024). Unlike supervised label-aligned topic models (Yang et al., 2025), our pipeline first induces generic locale-specific topics and only then maps topic centroids to Cultural Taxonomy leaves. This post hoc projection preserves the distinction between discovered corpus structure and the cultural audit categories used for comparison.

B Region Resolver Details

During lookup construction, each Common Crawl sample record is assigned a weighted vote ac-

cording to the agreement pattern among URL evidence, HTTP `Content-Language`, HTML `lang`, and language-identification metadata. The rules are used only during offline lookup construction; they do not fire independently during FineWeb/FineWeb-2 annotation.

Confidence-tier thresholds. After evidence aggregation, each URL-signature row is assigned a confidence tier based on the following thresholds. A row is labelled `high` if the winning region has at least 20 weighted votes, at least 90% vote share, and appears across at least three Common Crawl dumps. It is labelled `medium` if it has at least seven weighted votes, at least 75% vote share, and appears across at least two dumps. It is labelled `low` if it has at least three weighted votes and at least 70% vote share. These thresholds favour URL signatures that are stable across time and not merely frequent in a single crawl; we selected them empirically after manually investigating a few samples.

Real ccTLD evidence. When the URL contains a non-branded country-code top-level domain, the ccTLD supplies the primary region candidate. Three-way agreement among ccTLD, HTML `lang`, and HTTP `Content-Language` receives the highest weight. Agreement between the ccTLD and one header signal receives slightly lower weight. When the ccTLD is contradicted by a likely CDN-default header, the ccTLD is still trusted but down-weighted. When the ccTLD is contradicted by a country-specific header, the vote is retained with weaker weight. If no header signal is available, the ccTLD alone still contributes a strong but not maximal vote.

Path and subdomain evidence. For generic or branded domains, the system searches for locale-bearing path or subdomain tokens, such as `/uk/`, `/jp/`, `de.example.com`, or `en-gb`. These hints receive higher weight when corroborated by HTTP or HTML language metadata and lower weight when they appear alone.

Header-only evidence. If the URL contains no region cue, low-resolution header evidence may contribute a weak vote only when it is consistent with language identification. High-resolution headers on otherwise generic URLs are not treated as positive region evidence because such values are often injected by server or CDN defaults.

Fallback and unresolved records. If no reliable rule supplies evidence during lookup construction, the record contributes no vote. During search-time resolution, URLs that do not match the lookup table may still receive weak labels from curated URL hints or non-branded ccTLD fallback; otherwise the resolver emits `XX`.

Per-rule vote weights. Each rule contributes a per-record vote with a fixed weight, calibrated to reflect the strength of evidence it provides. The weights used during offline lookup construction are summarized in Table 5.

Rule	Evidence pattern	Weight
A1	ccTLD + HTML `lang` + Content-Language agree	3.0
A2	ccTLD + one header agree	2.5
A3	ccTLD beats high-resolution default	2.0
A4	ccTLD overrides country-specific header	1.0
A5	ccTLD alone (no header signal)	2.0
A6	Path/subdomain hint + header	2.0
A7	Path/subdomain hint alone	1.5
A8	Low-resolution header + language ID agree	1.5
A9	Soft header default on generic URL	0
A12	Unresolved	0

Table 5: Vote weights applied by each evidence rule during offline lookup-table construction. Higher weights correspond to stronger multi-signal agreement. Rules A9 and A12 are recorded for diagnostic completeness but do not contribute to weighted aggregation.

Lookup-Table Schema The frozen URL-signature lookup table is released as a CSV with 1,335,000 rows and the columns described in Table 6.

Column	Dtype	Meaning / values
`key_type`	str	`domain`, `host`, or `host_path`
`key`	str	URL signature (normalised, lowercased)
`n_docs`	int	raw record count contributing to the signature
`top_region`	str	winning region (ISO 3166-1 alpha-2) or empty
`share`	float	vote share of the winning region $\in [0, 1]$
`n_distinct_regions`	int	distinct candidate regions seen
`confidence`	str	`high`, `medium`, or `low`
`n_weighted`	float	total weighted votes for the winning region
`dump_diversity`	int	number of distinct CC-MAIN dumps the row appears in
`secondary_region`	str	runner-up region or empty
`secondary_share`	float	vote share of the runner-up
`n_content_votes`	int	A11 content-scan votes (0 in URL-only release)
`content_share`	float	A11 content-share $\in [0, 1]$ (0 in URL-only release)
`top_rule`	str	rule that contributed the largest mass (`A1`–`A12`)
`rule_mix`	str	semicolon-separated `Ak:count` list across all firing rules

Table 6: Schema of the released URL-signature lookup table.

Rows are emitted only when the winning region passes the confidence thresholds in Section 3.1.3. Null encodings: `top_region` and `secondary_region` use the empty string when no

candidate region is available; `n_content_votes` and `content_share` are zero for all rows in the URL-only release.

C search-Time Resolution Sources

At search time, the resolver applies the priority chain in Table 1. The first source that emits a valid region label determines the document's region, resolution source, and confidence value. The sources are interpreted as follows.

1. **Query parameters.** Explicit region parameters such as `gl=US`, `country=ca`, `region=GB`, or `locale=ja-JP` receive the highest priority. These cases are rare, but they are strong signals because the site operator is explicitly encoding a locale choice.

2. **Lookup-backed keys.** The next three sources consult the frozen lookup table. `host_path` captures multinational websites that organize regional sites under the first path segment, such as `pwc.com/uk`. `host` captures full hostnames such as `www.lemonde.fr`. `domain` collapses subdomains to a registered domain, allowing `news.bbc.co.uk` and `sport.bbc.co.uk` to inherit evidence from `bbc.co.uk` when appropriate.

3. **URL hints.** If the lookup table does not contain a matching key, the resolver searches for curated country tokens in structurally informative URL positions. Examples include `uk.airbnb.com`, `airbnb.com/jp/rooms`, and `redbubble.com/en-gb/`. Ambiguous tokens such as `co`, `it`, and `me` are excluded unless the URL structure makes the regional interpretation unambiguous.

4. **ccTLD fallback.** Finally, the resolver falls back to the URL's ccTLD when it is not on a branded-ccTLD denylist. The denylist includes ccTLDs widely used outside their nominal jurisdictions, such as `.tv`, `.ly`, `.io`, `.me`, `.ai`, `.co`, `.cm`, and `.gg`. Non-branded ccTLDs such as `.de`, `.fr`, `.jp`, `.ru`, and `.br` are retained as weak fallback evidence.

5. **No signal.** When all sources fail, the resolver emits `XX` with source `none` and confidence `none`. These records remain in the intermediate annotated corpus; filtering is left to downstream users.

D Topic-to-CT Softmax and JSD

Temperature-scaled softmax for topic-to-CT projection (Sec. 3.2). After topic induction, each merged topic is mapped to the CT label space. Each CT category is embedded once using BGE-M3 from its natural-language description. For each merged topic t and CT category c, we compute cosine similarity between the topic centroid e_t and the CT category embedding e_c, and convert similarities into a probability distribution over CT categories using a temperature-scaled softmax:

$$p(c \mid t) = \frac{\exp(\cos(e_t, e_c)/\tau)}{\sum_{c'} \exp(\cos(e_t, e_{c'})/\tau)}, \quad \tau = 0.05.$$

The low temperature makes the mapping sharp enough for the highest-scoring category to be interpretable, while still preserving non-zero mass on plausible runner-up categories.

Jensen–Shannon distance over LTDs (Sec. 5.3). For two locale-level Locale Topic Distributions P, Q over the 14 CT leaves, we report the Jensen–Shannon distance computed with $\log_2$:

$$\mathrm{JSD}(P, Q) = \sqrt{\tfrac{1}{2}\,\mathrm{KL}(P \,\|\, M) + \tfrac{1}{2}\,\mathrm{KL}(Q \,\|\, M)}, \qquad M = \tfrac{1}{2}(P + Q).$$

with KL in $\log_2$, so $\mathrm{JSD}(P, Q) \in [0, 1]$. Because per-locale leaf mass is produced by the temperature-scaled softmax above, every LTD has strictly positive entries on every leaf and no smoothing is applied; the KL terms are finite without an additive prior. Implementation: `scipy.spatial.distance.jensenshannon(p, q, base=2)`.

E Benchmark Audit Protocol

This appendix describes the construction and annotation protocol for the benchmark audit summarized in Section 4. The audit is designed to support corpus–benchmark comparison by coding each dataset along cultural coverage, scope, provenance, and usability dimensions.

E.1 Dataset Sources

We compile candidate datasets from three sources:

1. datasets included in the cultural-NLP taxonomy and survey of Liu et al. (2025);

2. datasets included in MUREL-style cultural-resource curation (Namazifard and Galke, 2025); and

3. newly identified datasets from systematic searches of the ACL Anthology, arXiv, HuggingFace, GitHub, and dataset repositories.

The systematic searches combine cultural terms with resource terms. Cultural terms include *culture*, *cultural*, *cross-cultural*, *multilingual*, *norms*, *values*, *bias*, *dialect*, *stereotype*, *social norms*, and *cultural knowledge*. Resource terms include *dataset*, *benchmark*, *corpus*, *evaluation*, and *test set*. Candidate records are deduplicated across source lists and benchmark families.

E.2 Inclusion and Exclusion Criteria

A resource is included if it satisfies all of the following criteria:

1. it provides a downloadable, inspectable, or otherwise accessible dataset;
2. it targets at least one cultural phenomenon covered by the CT, such as values, norms, dialects, communicative practices, culturally grounded knowledge, or culturally grounded bias;
3. it is documented in a paper, preprint, dataset card, or repository with sufficient metadata for annotation; and
4. it is usable as an evaluation dataset.

We exclude resources that do not provide dataset access, papers that discuss cultural evaluation without releasing or identifying a dataset, duplicate entries already represented by another variant, and resources whose cultural target cannot be mapped to any CT leaf.

When a benchmark family contains multiple variants, we count a variant separately only if it introduces a distinct language set, region set, task format, annotation schema, or culturally adapted item set. For example, a translated or culturally localised derivative is counted as a separate dataset but linked to its parent through the lineage field.

E.3 Annotation Dimensions

Each included dataset is annotated along four groups of dimensions.

Cultural coverage. These fields describe what cultural phenomena the dataset tests:

- CT branch;
- CT leaf or leaves;

Scope. These fields describe the populations, languages, and task format covered by the dataset:

- dataset name;
- geographical diversity;
- linguistic typology;
- languages covered;
- regions covered;
- dataset format;
- example item, when available.

Provenance. These fields describe where the dataset comes from and how it was constructed:

- paper link;
- publication venue;
- publication date;
- generation method: human, hybrid human–LLM, LLM-only, web-scraped or automatically extracted, or unclear;
- source list;
- dataset lineage.

Usability. These fields describe whether and how the dataset can be used for cultural evaluation:

- data link;
- accessibility status;

E.4 CT Leaf Annotation

CT branch and leaf assignments are coded from each dataset's paper, dataset card, repository, and example items. A dataset may receive multiple CT leaves if it evaluates more than one cultural phenomenon. For example, a benchmark containing culturally situated dialogue about politeness norms may be coded as both *Norms & Morals* and *Communicative Goals*; a multilingual stereotype benchmark may be coded as *Values–Bias*; and a dialect-identification benchmark is coded as *Dialects*.

Assignments are made at the dataset level rather than the individual item level. This choice allows broad survey coverage across the benchmark landscape, but it means that fine-grained within-dataset variation is not fully captured. Multi-category datasets are counted once per applicable CT leaf in the coverage matrix.

E.5 Lineage Coding

The lineage field records whether a dataset is original or derived from one or more parent datasets. We mark a dataset as derived if the paper or repository explicitly states that it is a translation, adaptation, localisation, extension, reannotation, or culturally modified version of an existing benchmark. Lineage is recorded using a parent-to-child notation, for example:

```
BBQ > MBBQ > KoBBQ
```

A dataset is coded as original if we find no known parent benchmark or source dataset beyond generic data sources such as surveys, web pages, or human-authored prompts. Lineage coding is conservative: if derivation is plausible but not documented, the dataset is not marked as derived.

E.6 Accessibility and Adaptation Coding

Accessibility is coded according to whether the dataset can be obtained and used for cultural evaluation:

- **Public**: the dataset is available through a repository, dataset hub, project page, or supplementary material.
- **Restricted**: the dataset requires an application, institutional access, or approval.
- **Unavailable**: the dataset link is broken, missing, or no longer accessible.

Adaptation feasibility records whether the dataset is directly usable as cultural assertions or evaluation items, or whether transformation is required. Common transformations include converting QA formats, extracting text from multimodal examples, converting classification labels into assertions, or filtering non-cultural items from broader multilingual datasets.

E.7 Quality Control

Factual metadata such as languages, regions, venue, publication date, dataset link, and format is extracted directly from source papers, dataset cards, or repositories. CT leaf assignments are coded by two annotators, with disagreements resolved through discussion. The final annotated table is released with dataset names, CT labels, language and region coverage, lineage fields, accessibility status, and annotation notes.

Artifact	Format	License
Region-annotated corpus	Parquet shards	ODC-By 1.0
URL-signature lookup table	CSV	CC-BY 4.0
Locale Topic Distributions	Parquet	CC-BY 4.0
Benchmark survey	BibTeX + CSV	CC-BY 4.0
Pipeline code	Git repository	Apache-2.0

Table 7: Released artifacts, formats, and licenses. The region-annotated corpus covers FineWeb and FineWeb-2 and inherits their ODC-By 1.0 license. Locale Topic Distributions cover 6,064 locales, and the benchmark survey contains 277 entries.

F Datasheet for FineWeb-CLaR

Following Gebru et al. (2021), we release a datasheet covering the six artefacts in our resource bundle. All artefacts share the same release URL and version-control history.

Motivation, composition, collection. The corpus is annotated from FineWeb (Penedo et al., 2024) and FineWeb-2 (Penedo et al., 2025); we do not redistribute the source text and reproduce only the metadata columns required for region attribution and topic-modeling.

PII and query-string filtering. The lookup table is keyed on `key_type` $\in$ {`domain`, `host`, `host_path`}; full URLs and query strings are never persisted in the lookup table. At search time, the resolver inspects query strings only to extract explicit locale parameters (`gl`, `country`, `region`, `locale`); other query parameters are discarded before annotation columns are written. No personal data, session tokens, or user identifiers are emitted in the released annotations.

Distribution, maintenance. Artefacts are released under the licences in Table 7 with a fixed version tag; future versions will be released alongside a changelog.

G Compute and Wall-Clock Disclosure

All region-attribution runs were single-machine, CPU-bound, with 64 worker processes against the frozen lookup table.

Topic modeling (BGE-M3 + FASTopic). Per-locale topic induction over 6,064 locales used multilingual BGE-M3 (Chen et al., 2024) for document encoding and FASTopic (Wu et al., 2024) for topic induction with $K = 200$ topics per locale. Wall-clock and GPU-hours were measured on 6$\times$ NVIDIA A100 80 GB GPUs in data-parallel configuration: Stage 0 (BGE-M3 embedding) used 65 GPU-h, Stage 1 (FASTopic induction) 48.0 GPU-h, and Stage 2 (CT projection + adjudication) 19 GPU-h, totaling $\approx$ 132 GPU-h; the realized cost is substantially lower than expected because the long tail of the locale corpus is dominated by much smaller locales.

LLM adjudication. Low-confidence topic-to-CT projections ($\max_c p(c \mid t) < 0.5$) were adjudicated by Qwen2.5-32B-Instruct (AWQ-quantised, served via vLLM 0.19.0 with `enforce_eager=True`) at temperature $T = 0$. Because the temperature-scaled softmax at $\tau = 0.05$ over the CT label space produces a flat maximum in practice (observed $\max_c p(c \mid t)$ in $[0.094, 0.253]$, mean 0.135), all of the 1,634 merged topics fell below the 0.5 cutoff and was adjudicated (i.e. 100% of merged topics triggered adjudication); the adjudicator changed the primary CT-leaf assignment for (31.6%) of them.

H Full Dataset-to-Taxonomy Mapping

Element	Datasets
Ideational	
Concepts (21)	Hu et al. (2024); Cao et al. (2023); Nayak et al. (2024); Romero et al. (2024); Winata et al. (2025); Myung et al. (2024b); Nguyen et al. (2023b); Guo et al. (2025); Majewska et al. (2022); Li et al. (2024d); Li et al. (2024e); Yin et al. (2021); Koto et al. (2024b); Liu et al. (2021); Hu et al. (2023); Wang et al. (2021); Ayash et al. (2025); Becattini et al. (2023); Thapliyal et al. (2022); Li et al. (2023b); Rei et al. (2023)
Knowledge (32)	Romero et al. (2024); Koto et al. (2024a); Myung et al. (2024b); Joy (2025); Nguyen et al. (2023b); Li et al. (2024c); Wibowo et al. (2024); Chiu et al. (2024b); Li et al. (2024d); Shi et al. (2024); Keleg and Magdy (2023); Hardalov et al. (2020); Semenov and Sennrich (2025); Yin et al. (2022); Singh et al. (2025); Koto et al. (2024b, 2023); Son et al. (2024); FitzGerald et al. (2022); Sakai et al. (2024); Lin et al. (2021); Kassner et al. (2021); Xia and Monti (2021); Pramodya et al. (2025); Jiang et al. (2020); Ponti et al. (2020); Yuan et al. (2024); Liu et al. (2024); Singh et al. (2024); Röttger et al. (2024); Lewis et al. (2020); Clark et al. (2020)
Values - general (15)	Scaria et al. (2024); Li et al. (2024b); Kirk et al. (2024); Aakanksha et al. (2024); Pistilli et al. (2024); Karinshak et al. (2024); Zahraei and Asgari (2025); Xu et al. (2024); Cahyawijaya et al. (2025); Li et al. (2024a); Durmus et al. (2024); Zhao et al. (2024); Sorensen et al. (2024); Al Kautsar et al. (2025); Hanges and Dickson (2004)
Values - bias (28)	Zulaika and Saralegi (2025); Joshi et al. (2025); Ravikiran and Annamalai (2021); Palta and Rudinger (2023); Nozza et al. (2021); Neplenbroek et al. (2024); Wang et al. (2025); Bhutani et al. (2024); Mitchell et al. (2025); Nadeem et al. (2021); Öztürk et al. (2023); Troles and Schmid (2021); Stanovsky et al. (2019); Kappl (2025); Xu et al. (2024); Lauscher et al. (2020); Parrish et al. (2022b); Nangia et al. (2020a); Strazda and Spanakis (2025); Névéol et al. (2022); Sahoo et al. (2024); Jin et al. (2023); Mukherjee et al. (2023a); Zhao et al. (2018); Lauscher and Glavaš (2019); Tomar et al. (2025); Rudinger et al. (2018); Röttger et al. (2024)
Values - hate (13)	Joshi et al. (2025); Ravikiran and Annamalai (2021); Mandl et al. (2021); Bassignana et al. (2018); Trager et al. (2025); Luu et al. (2021); Lee et al. (2023); Vargas et al. (2022, 2026); Mathew et al. (2020); Bui et al. (2025); Dementieva et al. (2025); Wulczyn et al. (2017)
Values - other perceptions (12)	Mohamed et al. (2022, 2024); Belay et al. (2025); Frenda et al. (2023); Havaldar et al. (2024); Deas et al. (2024); Pei et al. (2023); Casola et al. (2024); Yue et al. (2024); Barriere et al. (2025); Chakravarthi (2020); Hande et al. (2021)
Norms and Morals (36)	Chen et al. (2023); Guo et al. (2025); Wang et al. (2024); Joshi et al. (2025); Li et al. (2024b); Trager et al. (2025); Zhan et al. (2024, 2023); Xu et al. (2024); Cahyawijaya et al. (2025); Vargas et al. (2026); Huang and Yang (2023); Pyatkin et al. (2022); Yu et al. (2024); CH-Wang et al. (2023); Ohashi et al. (2024); Dwivedi et al. (2023, 2025); Hopp et al. (2021); Ziems et al. (2022b); Sahu et al. (2025); Hoover et al. (2020); Emelin et al. (2021); Rao et al. (2024); Ziems et al. (2023a); Li et al. (2023a); Fung et al. (2023); Kim et al. (2022); Saffari et al. (2024); Yuan et al. (2024); Forbes et al. (2020); Hobson et al. (2024); Sorensen et al. (2024); Stranisci et al. (2021); Al Kautsar et al. (2025)
Artifacts (12)	Mohamed et al. (2022, 2024); Koto et al. (2024b); Hossain et al. (2022); Tikhonov et al. (2021); Becattini et al. (2023); Hobson et al. (2024); Epure et al. (2020); Jiang et al. (2023); Al Kautsar et al. (2025); Neudecker (2016); Dhanwal et al. (2020)
Linguistic Elements	
Dialects (15)	Winata et al. (2025); Wibowo et al. (2024); Frenda et al. (2023); Singh et al. (2025); Casola et al. (2024); Dementieva et al. (2025); Ziems et al. (2023b); Kuparinen et al. (2023); Faisal et al. (2024); Olabisi et al. (2022); Abdul-Mageed et al. (2023); Plüss et al. (2022, 2023); Ziems et al. (2022a); Singh et al. (2024)
Styles/Reg./Genres (16)	Winata et al. (2025); Guo et al. (2025); Sweed and Shahaf (2021); Nadejde et al. (2022); Krishna et al. (2022); Khoshtab et al. (2025); FitzGerald et al. (2022); Zhan et al. (2023); Srinivasan and Choi (2022); Briakou et al. (2021); Li et al. (2023b); Laippala et al. (2022); Sun and Xu (2022); Kuzman et al. (2023); Havaldar et al. (2023); Mukherjee et al. (2023b)
Social Elements	
Relationship (7)	Wang et al. (2024); Chiu et al. (2024b); Zhan et al. (2024, 2023); Li et al. (2023a); Shaikh et al. (2023); Hanges and Dickson (2004)
Context (17)	Wang et al. (2024); Ravikiran and Annamalai (2021); Wang et al. (2020); FitzGerald et al. (2022); Currey et al. (2022); Wang et al. (2021); Pramodya et al. (2025); Goel et al. (2023); Zhan et al. (2024, 2023); CH-Wang et al. (2023); Emelin et al. (2021); Ziems et al. (2023a); Kim et al. (2022); Forbes et al. (2020); Laippala et al. (2022); Kuzman et al. (2023)
Communicative Goals (11)	Wang et al. (2024); Goel et al. (2023); Zhan et al. (2024, 2023); Emelin et al. (2021); Ziems et al. (2023a); Li et al. (2023a); Laippala et al. (2022); Kuzman et al. (2023); Jin et al. (2024); Hanges and Dickson (2004)
Demographics (12)	in the Wild (2024); Frenda et al. (2023); Trager et al. (2025); Pei et al. (2023); Casola et al. (2024); Kirk et al. (2024); Lee et al. (2023); Ziems et al. (2023a); Saffari et al. (2024); Zhao et al. (2024); Shaikh et al. (2023); Plüss et al. (2023)

Table 8: Datasets organized by taxonomy element, following Liu et al. (2025). Numbers in parentheses indicate dataset count per category. Multi-category datasets appear under each applicable element.