How Good is Your Wikipedia? Auditing Data Quality for Low-resource and Multilingual NLP

Kushal Tatariya[1*] **Artur Kulmizev**[2*] **Wessel Poelman**[1] **Esther Ploeger**[3]
Marcel Bollmann[4] **Johannes Bjerva**[3] **Jiaming Luo**[5] **Heather Lent**[3] **Miryam de Lhoneux**[1]
[1]Department of Computer Science, KU Leuven [2]CENTAL, UCLouvain
[3]Department of Computer Science, Aalborg University
[4]Department of Computer and Information Science, Linköping University
[5]Google Translate

Abstract

Wikipedia's perceived high quality and broad language coverage have established it as a fundamental resource in NLP. However, in recent years, such assumptions of high quality have become the subject of scrutiny in low-resource and multilingual contexts. In this study, we subject the entirety of non-English Wikipedia to a data filtering procedure typically reserved for noisy web-text — a process which removes a large percentage of the collection's data. In analyzing the removed data, we reveal numerous systematic quality issues, such as script and language contamination, repeated template and placeholder articles, and a high concentration of bot-generated content. We consolidate these findings into a 4-level quality ranking of Wikipedia, which shows strong correspondence with alternative quality measures and heuristics. Lastly, we evaluate the downstream impact of quality filtering in three practical language modeling scenarios, showing that models trained on filtered data largely match or outperform those trained on raw Wikipedia, with the largest gains observed for lower-quality language editions. Ultimately, our experiments serve as a first step in establishing quality-aware best practices for Wikipedia utilization in NLP, laying groundwork that can inform future dataset creation and curation efforts.

1 Introduction

Wikipedia has long served as an invaluable NLP resource, representing an international collaborative effort in documenting an ever-changing world. Its open-source framework — through which authors, editors, and administrators cooperate to continuously refine and expand content — ensures that information contained therein is not only current, but also well-written, accurate, and relevant to any given subject. For these reasons, Wikipedia has been employed for a myriad of NLP applications, such as language model (LM) training and benchmarking (Merity et al., 2016), question-answering (Hewlett et al., 2016; Rajpurkar et al., 2016; Joshi et al., 2017) and knowledge-base creation (Vrandečić and Krötzsch, 2014; Lehmann et al., 2024), among many others. Wikipedia is also a staple resource in multilingual NLP, where it is commonly employed for pretraining multilingual LMs (Devlin et al., 2019; Conneau and Lample, 2019) and creating benchmarks for downstream tasks such as named-entity recognition (Rahimi et al., 2019) and machine translation (Schwenk et al., 2021).

Recently, increasing attention to data quality has led to the adoption of various commonplace data filtering practices within NLP, whereby noisy web-scale data is distilled into a more salient signal for LM pretraining. Wikipedia has often been employed as a 'high quality' reference in this process (*e.g.,* Wenzek et al., 2020), largely due to its encyclopaedic domain and culture of rigorous community moderation. However, the assumption of Wikipedia's inherent 'high quality' — though generally reliable for English and select high-resource languages — is now increasingly being challenged in the multilingual setting (Kreutzer et al., 2022). Indeed, it has been found that Wikipedias for certain lower-resourced languages can contain unnatural, inorganic, and machine-translated text, among other noise, which can be incomprehensible to native speakers (Alabi et al., 2020; Lent et al., 2024). As a result, poor data quality for such languages typically leads to poor performance and untrustworthy models, ultimately providing unusable technologies for speakers (Held et al., 2023; Nicholas and Bhatia, 2023; Durmus et al., 2024).

In this study, we critically examine the extent of Wikipedia's 'high quality' in a non-English setting. Estimating data quality via linguistic analyses and native speaker judgments is a costly endeavour for individual Wikipedias, and not feasible across the

[*]Equal contribution

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29754–29774

collection's 340+ active languages with an estimated 65 million articles. Thus, we operationalise existing automatic data cleaning methods to reliably determine Wikipedia quality in a data-driven manner. Specifically, we employ two distinct data cleaning processes — *script filtering* and *deduplication* — and apply them successively to every available Wikipedia, excluding English. Provided that these techniques are typically applied to noisy, uncurated web-text, the extent to which a given Wikipedia is robust to such operations serves as an inverse measure of its 'quality'. We demonstrate that each cleaning step is able to target distinct categories of noise, revealing systematic differences in content quality across language editions — regardless of size. Furthermore, we find that the downstream performance of models trained on 'cleaned' Wikipedias is often preserved — and sometimes improved — with respect to the original, unfiltered data. Overall, our contributions are as follows:

1. We present empirical evidence of quality issues across Wikipedia, establishing a quality taxonomy that can be employed for future research.
2. We perform downstream evaluation of filtered Wikipedias across three practical low-resource and multilingual scenarios.
3. We release an open-source NLP toolkit for cleaning text data in many languages.[1]

2 Related Work

2.1 Data Filtering

Recent studies in large language modelling have demonstrated the benefits of filtering web-scale data (Wenzek et al., 2020; Raffel et al., 2023; Penedo et al., 2024; 01.AI et al., 2024). In this context, data filtering typically involves a combination of techniques, ranging from basic document-level heuristics to corpus-level statistics that serve as quality proxies (Albalak et al., 2024). For example, the popular C4 dataset includes heuristics that remove documents containing code (*e.g.*, 'Javascript'), boiler plate legal policies (*e.g.*, 'Terms and Conditions'), or inappropriate terms, among other hand-crafted rules (Raffel et al., 2023). Alongside heuristics, *deduplication* is another common filtering technique designed to remove duplicate documents from a given dataset, only keeping one copy thereof. A particularly popular algorithm is MinHash, due to its effectiveness in removing semantically identical documents, as well as its scalability across CPU nodes (Penedo et al., 2024). MinHash has likewise been combined with downstream clustering-based deduplication to filter documents in a more aggressive manner (Tirumala et al., 2023). More sophisticated techniques have recently also been employed: for example, 01.AI et al. (2024) leverage learned filters (*i.e.*, quality classifiers and cluster-based filters), which are designed to preserve documents resembling a 'high quality' reference corpus – typically Wikipedia (Wenzek et al., 2020; Together Computer, 2023).

2.2 Wikipedia in Multilingual NLP

Wikipedia's widespread language coverage and reputation as a high-quality data trove has made it a vital resource in multilingual NLP applications. It has served as the backbone of many multilingual benchmark datasets, such as WikiAnn (Pan et al., 2017), XQuAD (Artetxe et al., 2020), MLQA (Lewis et al., 2020), TyDiQA-GoldP (Clark et al., 2020), and BUCC (Zweigenbaum et al., 2017), as well as a common source of pretraining data for popular multilingual LMs, such as mBERT (Devlin et al., 2019) and XLM (Lample and Conneau, 2019). Beyond this, it was notably employed as the training data for CCNet's language identification classifier, which is primarily used for partitioning the popular Common Crawl dataset along language boundaries. Given that Common Crawl serves as the basis for many contemporary open-source pretraining datasets, the quality and performance of such a classifier becomes paramount in retaining relevant data.

As a consequence of its ubiquity in NLP, Wikipedia has recently attracted increased scrutiny for its data quality. For example, Kreutzer et al. (2022) find that the Wikipedia-oriented WikiMatrix (Schwenk et al., 2021) contains a number of deprecated, incorrect, or otherwise mislabelled language codes, among the 86 languages included (Kreutzer et al., 2022). Lignos et al. (2022) identify similar shortcomings for lower-resourced languages in WikiAnn and even Wikidata.[2] Studies on selected low-resource languages likewise confirm such findings. For example, Alabi et al. (2020) observe that the Yorùbá and Twi Wikipedias are ostensibly written by non-native speakers and lack the necessary diacritics and variation to be relevant

[1] https://github.com/akulmizev/texieve

[2] https://wikidata.org

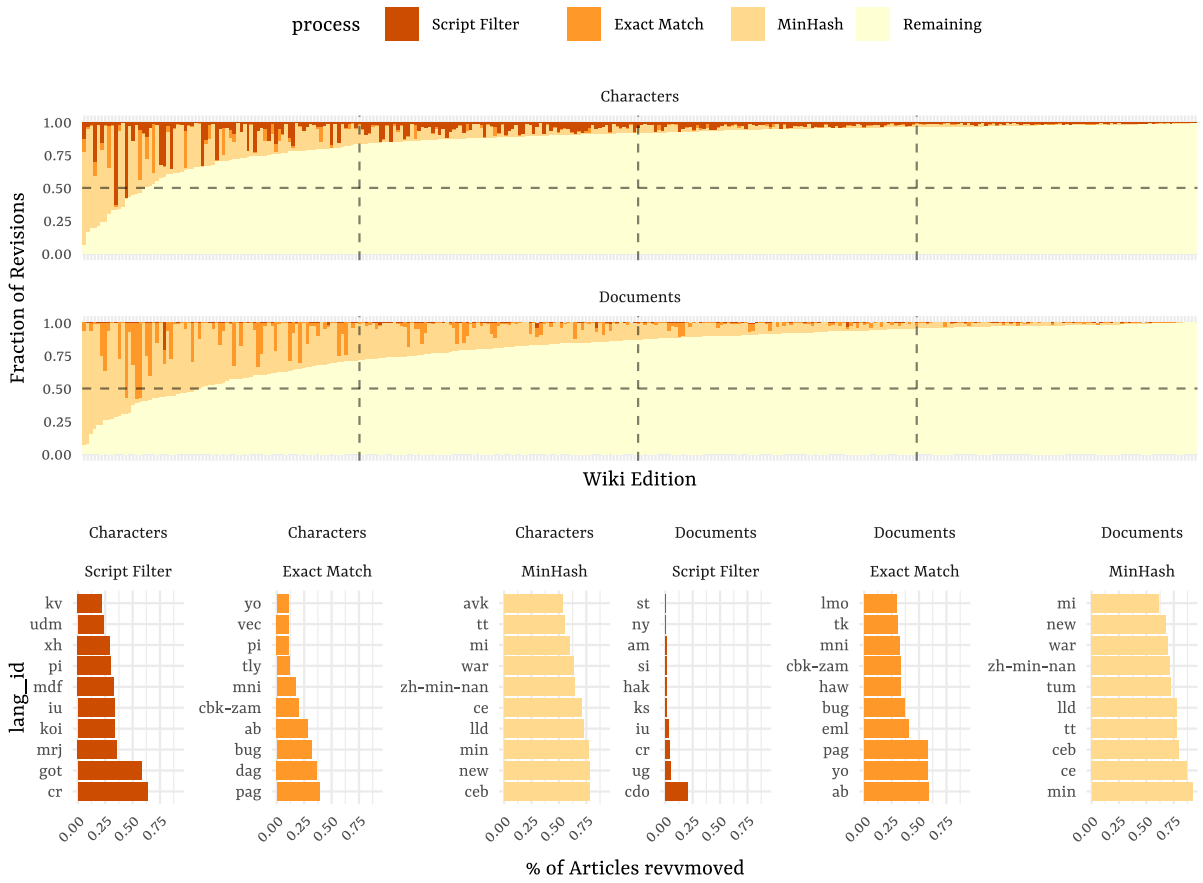

Figure 1: Top: Fraction of data removed by each filter, sorted by remaining data in each Wikipedia. Bottom: Top 10 Wikipedias according to fraction of data removed by each filter.

to speakers, while Lent et al. (2024) find that sentences from Wikipedias for Creole languages are often incomprehensible by speakers and require manual correction.

2.3 Data Quality versus Data Quantity

The notion of data *quality* is inherently subjective. Quantitatively, it is commonly measured by performance gains on a downstream NLP task. Qualitatively, it can be understood to signify utterances that are accepted as valid by proficient speakers of a language, or, alternatively, text samples that do not promote toxic or offensive sentiments (Kreutzer et al., 2022; van Noord et al., 2024). Such definitions can sometimes be at odds with one another. For example, Longpre et al. (2024) show that a smaller, carefully filtered dataset can improve performance on downstream tasks in contrast to a larger, unfiltered dataset, but that the resulting model becomes more likely to generate toxic content. Likewise, diminishing returns have been demonstrated from training on ever-increasing amounts of data (Warstadt et al., 2023). For example, language models pretrained on C4 (745GB) marginally outperform those trained on Wikipedia (16GB) over benchmarks like GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and SQuAD (Rajpurkar et al., 2016). However, many of these benchmarks were sourced from Wikipedia itself, suggesting that quantity is less important than domain (Raffel et al., 2023).

3 Is Wikipedia 'High Quality'?

Albalak et al. (2024) outline a set of best practices for quality data selection and filtering. In this work, we examine the effects of their proposed model-agnostic filtering strategies,[3] as applied to Wikipedia. Conceptually, we can recognize such filters as belonging to two distinct categories based on their underlying objective: 1) procedures for removing undesirable material from Wikipedia, such as foreign scripts and duplicate articles; and 2) heuristics that aim to enhance the existing data dis-

[3]We exclude classifier-based and perplexity-based filtering strategies measuring quality relative to a selection of data.

Process	Text	Source Title (Link)
Script Filtering	EΘ ᏉᎵ ᏫZᏢᏞ ᎦhᏓᏣ,DᏖW ᎤᏁᎦ [lining that extends from the tip of its wing along the rear of the wing to its body, Zᏼ a head with reddish skin. The EhᎻ ᏉᎵ has a shorter tail, silver splotches near the tips of its wings, Zᏼ a head with grayish skin. Both EhᎻ ᏉᎵ Zᏼ EΘ ᏉᎵ ᏯᏆᏯᎽ ᏂᏯᏆ locate the dead animals or carrion they feed upon by both sight Zᏼ smell.	ᏂᏯᏆ
	DhGWᎽ DhᏫhᏯᎽ polysynthetic ᎦᏫhᏯᏯᏗ. GWᎽ ᎦᏫhᏯᏯᏗ OᏂᎻᏯ ᏰᏓᏁW Iroquoian ᏧhᏫhᏯᏯᏗ, D4Z, ᏯᏢᏯᏓ ᏗᏆᎦᏖᏰR, ᎻᏢᏢ ᎦᏫhᏯᏯᏗ ᏯᎽ. DhGWᎽ DhBΘᏯ ᏧZᏯG, ᎤᏁWOᏂᏯ ᎦᏫhᏯᏯᏗ OᏫᏮ ᏂᏢᏢRT. 1820ᏼ, ᏯᏰᏢᏯ, Dᏼ George Gist, ᏯD GWᎽ ᏗAᏨᏢ AᏨᏢᏯᏂᏢT. ZᎽᏢ ᏯD Unicode TᎽᏂ (ᏂΘΘᏗP). ᏯDZ ᎦᏓᏬ, RᏢᎽᏢ, GWᎽ TAᏨᏀᏗ ᎤᏫᏂᏒᏓ ᏫᎾᎦᏢᏯᎽ. ᏯDᏯᏯ ᏎᏓT ᏮᏂ ᎤᏚᏢᏫᏢ hᎦᏢᏯᏗᏢT: 2007 ᏂᏢᏢR, iᏂBᏢ ᏗᏚᏫᎽ (Myrtle Driver Johnson) DEᏯ TᏀᏫᏬW 1800Ꮟ ᏂᏢᏢR, Charles Frazier ᏗAᏨᏢ ᎤᏉᏙᏢ, KᎦS OᏉV: ᏂᏢEᏫᎥᏗᏯᏢT GWᎽ DᏁᏯᏗᏯᎽ AᏨᏢᏯᏯᏢT.	ᎶWᎽ ᏫᏫhᏯᏯᏗ
	ᏭᏀA (Ma-lo-go), ᏉᏓᎽ ᎤᎾLCᏂᏯ ᎤEΘᏀᏯ ᏭᏀA (Al-Mamlakah al-Maɣribiyya), DᏚᏢ ᎽᏎᎦᎽ DGᏯ.	ᎤEΘᏀᏯ
Exact Match	Tiŋ ŋɔ nyɛla tiŋ' shɛli din be Zabzugu Municipal Assembly, Northern Region, Ghana tiŋ gbaŋ ni.	Sawieb
	Tiŋ ŋɔ nyɛla tiŋ' shɛli din be Zabzugu Municipal Assembly, Northern Region, Ghana tiŋ gbaŋ ni.	Tayondo
	Tiŋ ŋɔ nyɛla tiŋ' shɛli din be Zabzugu Municipal Assembly, Northern Region, Ghana tiŋ gbaŋ ni.	Poagmado Polinaayili
MinHash	Say Central Visayas et rehiyon na Filipinas. Unong ed 1 Hulyo 2024 census, say populasyon to et 6,640,875 totoo tan 1,966,588 abong. Walay kabaleg tan sukat to ya 15,895.66 sq. km.	Central Visayas
	Say Ilocos Region et rehiyon na Filipinas. Unong ed 1 Hulyo 2024 census, say populasyon to et 5,342,453 totoo tan 1,306,256 abong. Walay kabaleg tan sukat to ya 13,012.60 sq. km.	Ilocos Region
	Say Central Luzon et rehiyon na Filipinas. Unong ed 1 Hulyo 2024 census, say populasyon to et 12,989,074 totoo tan 2,511,783 abong. Walay kabaleg tan sukat to ya 22,014.63 sq. km.	Central Luzon

Table 1: Examples of Wikipedia articles targeted by each filtering process. The selected languages are: Cherokee (`chr`, top), Dagbani (`dag`, middle), and Pangasinan (`pag`, bottom).

tribution by removing otherwise lower quality text according to a given metric (e.g. article length in characters).

We apply select filters from the first category to identify undesirable content in non-English Wikipedias — namely script filtering and deduplication. We posit that if these filters remove a substantial amount of text from a given Wikipedia dump, that content was not of high quality to begin with. Ultimately, we find that 11.68% and 29.36% of total characters and documents are removed from the raw Wikipedia, respectively, as shown in Figure 1 (top). In what follows, we will analyse each filtering step, as well as the languages most affected by them.

Script Filtering Applying an existing language identification model to Wikipedia is a challenge due to data contamination – most models supporting all of Wikipedia's languages have generally also been trained on Wikipedia (see e.g. Kargaran et al., 2023). Thus, we employ script-based filtering as a proxy for language identification by filtering characters in the Wikipedia that are not in the officially documented script(s)[4] for the concerned language (see Appendix A for more details).[5] We find that 3.33% of characters in non-English Wikipedias are removed by this process.[6] Interestingly, the relative amount of removed documents is much smaller at 0.01% (7782 total), which indicates that this text stems from very long, potentially copied and pasted articles. For example, consider the article for Alexander Fleming in Assamese, which is mostly a direct copy of the corresponding English version. This particular article is a symptom of the problems associated with the use of machine translation tools[7] in Wikipedia content creation, which can lead to entire sections of articles being mistranslated, or not translated at all (Wilson, 2019). With this in mind, script filtering can utilized as a crude way to remove mistranslated content. However, the problem of detecting translationese in Wikipedia is complex and well documented (Judah, 2025) — especially for low-resource languages — and is not easy to mitigate without human evaluation.

Outside of translation artefacts, it is important to note that the vast majority of characters removed from script filtering stem from low-resource or extinct languages (for an example, refer to Table 1

[4]For example, the Konkani Wikipedia features 3 scripts – Devanagari, Latin and Kannada.

[5]https://en.wikipedia.org/wiki/List_of_Wikipedias

[6]We acknowledge that this methodology would not be effective for English articles in non-English Wikipedias written in the Latin script. However, this would require a language ID model that has not been trained on Wikipedia. We also attempted to use a language ID model to exclusively classify and filter English text, but we found that this led to over-filtering and misclassification for some Creoles and Pidgins, eg. Naija.

[7]e.g. https://www.mediawiki.org/wiki/Content_translation

(top)), such as Cree (cr, 187 documents), Cherokee (chr, 1113), or Gothic (got, 1013). In most cases, these are Latin characters corresponding to placeholder information, transliterated named entities or embedded links (if improperly formatted). In many cases, Wikipedia is among the few public resources that collects text for such languages, making it crucial to apply filters that preserve authentic, language-specific content — especially before downstream use in dataset creation, tokenizer training, or other applications.

Exact-match Deduplication We employ a hashing function to identify and remove articles which have the exact same text, but a different unique ID or title. This removes 0.16% and 1% of all characters and documents, respectively. However, we find that some Wikipedias are disproportionately affected by this process. For example, the Yorùbá Wikipedia (yo, 33,819 documents) loses almost 60% of its documents — many of which contain a single token *Ìtọ́kasí* (Reference). In general, articles typically removed by exact-match deduplication are placeholders containing filler or templatic text and generally contain little usable information. As an illustration of this, refer to Table 1 (middle), which shows a template assigned to 48 distinct municipality pages in the Dagbani (dag) Wikipedia. Accounting for such phenomena — especially at the document level — ensures that resource statistics are not artificially inflated and accurately represent the amount of authentic data available in a given language.

MinHash Deduplication While exact-match is able to filter identical documents, it cannot account for partial duplicates — *i.e.*, sets of documents containing mostly identical text, but varying in terms of single sentences, paragraphs, or even hyperlinks. For example, the Yorùbá articles for the Baluch people and the Baluchi language are largely identical, with the latter containing an additional language data panel (see Table 1 (bottom) for another example in Pangasinan). To capture such cases, we perform *MinHash deduplication* with locality sensitive hashing (LSH) and discount any article that returns a Jaccard similarity over 0.85. We find that a staggering 28.33% of all non-English Wikipedia articles (and 8.18% characters) are affected by this procedure, largely consisting of Wikipedias known to be primarily bot generated (more in §4.3). Here, it is important to note that many entries with high MinHash removal rates are likewise among the largest in terms of total document count prior to filtering. These include Cebuano (ceb, 6.1 million documents, 80% removed), Swedish (sv, 2.6, 47%)[8], Dutch (nl, 2.1, 35%), and Vietnamese (vi, 1.3, 58%) — the raw data of which was ostensibly used to train the highly influential multilingual BERT model.[9]

4 Characterizing Quality

Having demonstrated the effect of filtering, we now attempt to shed light on how our approach relates to a generalized notion of dataset quality (as it pertains to Wikipedia). To do so, we first categorize all non-English Wikipedias into 4 tiers — Tier 1, 2, 3 and 4 — based on how much data is removed during filtering (ranked from least to most). We perform k-means clustering, where each Wikipedia is represented by the percentage of remaining documents and characters. Based on several iterations of the clustering algorithm and manual inspections, we arrive at the ranking illustrated in Figure 2 (left). A list of all Wikipedias and their respective quality tiers can be found in Table 15.

4.1 Resource Availability

At first glance, our quality ranking might seem to categorize Wikipedias based on resource availability. To assess how accurate this impression is, we compare our classification with the taxonomy proposed by Joshi et al. (2020), which divides the world's languages into 6 categories according to the volume of labelled and unlabelled data available for them (with the latter measured by Wikipedia document counts). Figure 2 (right) presents a comparison between their taxonomy and our quality tiers. As expected, we observe that Joshi et al. (2020)'s categorization has a strong correlation with Wikipedia size, but demonstrates weak alignment with our quality taxonomy, suggesting that quality is not simply a direct consequence of resource availability.

While we observe that most Winners unsurprisingly are allocated to Tier 1, Arabic falls to Tier 2. Upon a cursory inspection, we find that 15% of the original Arabic Wikipedia is removed via MinHash deduplication, indicating that it contains a sizeable volume of potentially templatic or bot-generated text. We notice a similar effect for other

[8] see Lokhov (2021)

[9] https://github.com/google-research/bert/blob/master/multilingual.md

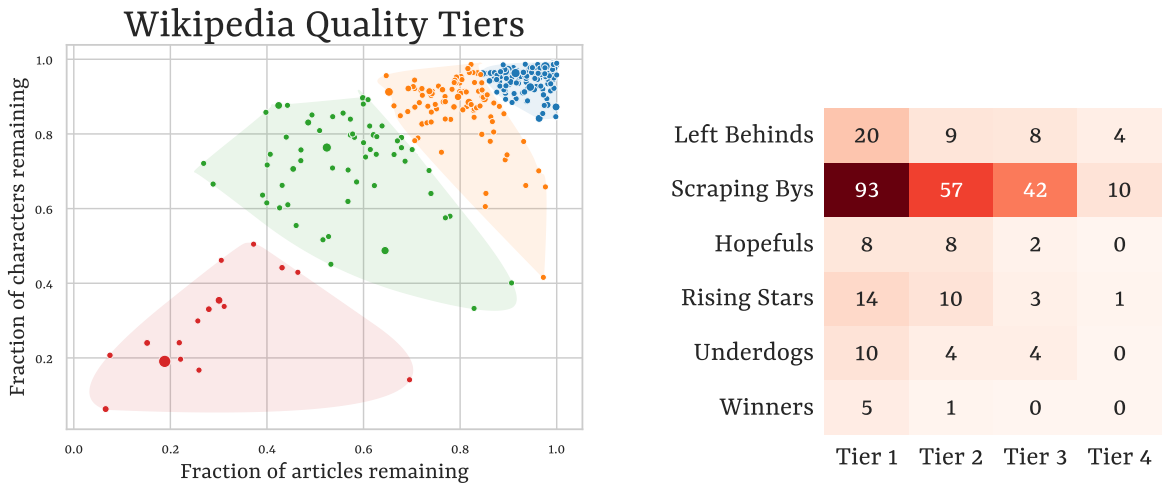

Figure 2: *Left*: Quality clusters based on the fraction of documents and characters removed after filtering, respectively. Clusters correspond to Tier 1 (blue), Tier 2 (orange), Tier 3 (green) and Tier 4 (red). Each point is weighed by each Wikipedia's unfiltered document count. *Right*: A confusion matrix comparing the Joshi et al. (2020) and the Wikipedia quality tiers.

higher-resourced Underdogs and Rising Stars languages, which are placed in Tier 3 (Swedish, Vietnamese, Serbian and Basque) and Tier 4 (Cebuano) — also due to their vulnerability to MinHash (as noted in §3). Going further, we find that 113 of the lowest resourced languages classified as Scraping Bys and Left Behinds (35% of Wikipedia in total) are likewise categorized as having Tier 1 Wikipedias, implying the involvement of dedicated moderation communities. Conversely, we also observe many Scraping Bys and Left Behinds languages falling into the Tier 3 and Tier 4 categories, such as Yorùbá and Twi, indicating that they contain serious quality issues that can be harmful to low-resource NLP.

4.2 Comparison to Depth+

In its official statistics, Wikimedia provides a Depth metric (alternatively, *editing depth*), designed to serve as a proxy measure for a given Wikipedia's 'collaborative quality'.[10] However, Alshahrani et al. (2023) caution against utilizing it for such purposes. Specifically, they discuss the metric's sensitivity to bot content (articles and edits) and excessive editing ('edit wars'), as well as its inability to consolidate broader user activity and account for lower-resourced Wikipedias. As such, they propose a modified metric, Depth+:

$$\text{Depth}^{+} = \text{Editors} \cdot \frac{\text{Edits}}{\text{Total}} \cdot \frac{\text{Articles}}{\text{Non} - \text{Articles}} \quad (1)$$

[10] https://meta.wikimedia.org/wiki/Wikipedia_article_depth

Alshahrani et al. (2023) show that Depth+ is able to adequately account for bot-generated content, correctly de-ranking Wikipedias such as Vietnamese, Serbo-Croatian, and Cebuano. Conversely, they demonstrate that Depth+ likewise includes several low-resource Wikipedias, such as Cree, Tigre, and Bangla towards the top of the ranking, highlighting their users' supposedly high degree of collaboration.

We find a significant positive correlation between our filtering procedure (measured as the ratio of percentage of documents and characters retained) with Depth+ ($\rho = 0.31, p < 0.001$). Thus, both metrics appear to capture similar notions of quality. However, approximately half of all Wikipedias yield Depth+ values of less than 0.3, and a majority very close to 0. The rest yield significantly higher Depth+ values (excluding English, German is highest with 36.04) resulting in many outliers that skew the distribution. We surmise that this is because the number of edits is disparagingly different across Wikipedia, effectively privileging Wikipedias where active users comprise a larger percentage of total users — a proxy for 'collaborative quality'. However, this does not shed much insight on the quality of individual articles written by one-time, potentially domain-expert authors.

4.3 Bot generation

We have observed that our filters remove a sizeable volume of articles from Wikipedias known to contain proliferate bot-generated content. To better un-

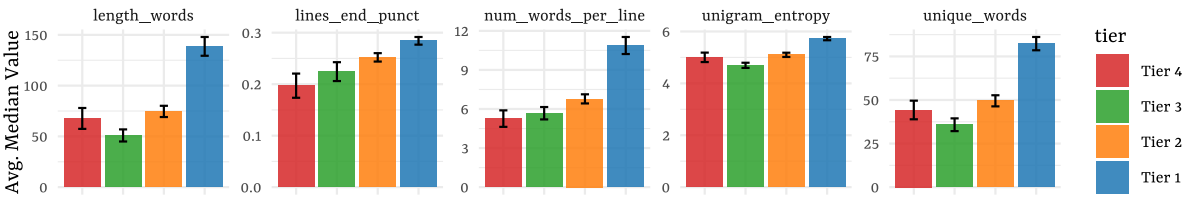

Figure 3: Aggregate median heuristic values by Tier. Spread across Wikipedias indicated by error bars.

derstand if we are indeed filtering out such content, we use Alshahrani et al. (2023)'s estimation of bot-generated articles across Wikipedia, which draws from user registration information and other metadata. Indeed, we find that there is a strong negative correlation ($\rho = -0.56, p < 0.001$) between the percentage of bot articles in a given Wikipedia and the percentage of articles remaining after filtering. We consider this to be an encouraging finding, provided that our data filters are simple, pre-defined processes that state obvious assumptions about our ideal data distribution. Additionally, we do not rely on specialized bot-detection models trained on other distributions.

We further hypothesize that MinHash deduplication plays a large role in removing bot articles, given its ability to fuzzy-match strings across documents. Here, we observe that the percentage of documents filtered out by MinHash yields an even higher correlation with bot article ratio ($\rho = 0.63, p < 0.001$), indicating that this process is an adequate proxy for bot article removal in Wikipedia. Unsurprisingly, the Wikipedias with the highest percentage of bot content according to Alshahrani et al. (2023) — Cebuano (99%), Waray (90%), Swedish (68%) — tend to have much (if not most) of their content removed by MinHash.

4.4 Heuristics as Quality Measures

Albalak et al. (2024) define *heuristic filtering* as a process that computes document-level statistics to eliminate low-quality data that falls below a specified threshold. Heuristics, in this sense, refer to simple metrics that aim to numerically characterise document content — for example, the number of characters contained in a given document. Though we do not attempt to filter Wikipedia according to these metrics (see Appendix B), we are nonetheless interested in how their distributions align with our quality tiers. In order to investigate this, we calculate five common and weakly correlated heuristics, as implemented in the Red Pajama (Together Computer, 2023) dataset cleaning pipeline: length in words, number of unique words, unigram entropy, average number of words per line, and fraction of lines ending in punctuation.[11] Due to the fact that many heuristics tend to follow an exponential distribution, we calculate each metric at the document level and record the median value for each Wikipedia (refer to Appendix E for each median heuristic value per edition, per tier).

In averaging median heuristics across Wikipedias, we observe a clear, monotonic pattern that neatly matches our ranking: Tier 1 produces the best values (in terms of polarity), followed by Tier 2 and then Tier 3 (Figure 3). This is largely intuitive: high quality articles can be assumed to be long (`length_words`), verbose (`unique_words`), less lexically predictable (`unigram_entropy`), as well as to contain longer (`num_words_per_line`) and better-formatted paragraphs (`lines_end_punct`). The only consistent exception is that of Tier 4, which does not follow monotonicity for 3 out of 5 heuristics, landing in between Tiers 2 and 4. However, this is largely expected: Tier 4 contains Wikipedias that are especially vulnerable to select filtering steps, which can affect heuristic measurement in different ways. For example, Cebuano scores high on `lines_end_punct` due to being largely bot generated, but likewise yields low `unigram_entropy` values for the same reason. Indeed, we can confirm this to be the case by observing the spread of median heuristic values, with Tier 4 being the highest for each metric.

5 Downstream Evaluation

The most salient use case for Wikipedia within NLP has been language modelling. This section examines how low-quality Wikipedia content affects this domain across three different contexts: train-

[11] The remaining heuristics are largely tailored to web text and are thus not directly relevant to Wikipedia, such as mentions of "javascript" or "lorem ipsum", or excessive within-document string duplication.

Figure 4: *Left*: Performance difference over SIB200 after monolingual pretraining on **raw_wiki** and **+filters** data; positive values indicate preference for the latter. Colours correspond to quality tiers described in §4 *Center*: KDE plots of SIB200 score distributions of **raw_wiki**, **+filters**, and **raw_wiki** models, organized by tier. *Right*: Performance difference over SIB200 after language adaptation. All scores are an average of 5 runs. Full result tables for all tasks are in Appendix C.

ing models in a single language, adapting models to new languages, and training multilingual models. This analysis compares two datasets: the complete, unfiltered Wikipedia data (**raw_wiki**) versus data cleaned using the methods outlined in §3 (**+filters**).[12] The underlying hypothesis is straightforward — if the performance difference (Δ) between models trained on **+filters** and **raw_wiki** is negligible, this would confirm that the removed data was indeed low quality.

5.1 Monolingual Pretraining

In order to minimize confounding factors such as language imbalance or poorly calibrated tokenizers, we pretrain mini monolingual DeBERTa (He et al., 2021) models (approximately 10 million parameters) on **raw_wiki** and **+filters** data from 50 languages and evaluate them on topic classification with SIB200 (Adelani et al., 2024).[13]

The left panel of Figure 4 shows Δ for each Wikipedia edition relative to the proportion of retained data. Here, we observe that most values fall within the $(-0.05, 0.05)$ range, suggesting that the removed content is indeed low quality and does not contribute meaningfully to model training. A notable example is Cebuano (`ceb`), which achieves a 12% performance improvement with **+filters** — a gain we largely attribute to the deduplication process eliminating most bot-generated content. In contrast, Acehnese (`ace`) appears negatively impacted by filtering, with $\Delta = -0.12$. Closer examination reveals that while this Wikipedia is similarly affected by MinHash deduplication (losing 42% of characters and 54% of documents), it contains far fewer raw documents than Cebuano — only 13,000 compared to 6 million. This suggests that such substantial data reduction can hurt performance in low-resource scenarios, even when the removed content resembles bot-generated content.

In order to confirm that **+filters** indeed removes low quality documents, we conduct an additional control experiment (**random**), wherein we sample n random articles from **raw_wiki** to match the size of **+filters**. Figure 4 (center) shows the distribution of scores for all three conditions separated by tier. Here, we observe that the distribution of the **random** scores is, on average, substantially lower than the other methods, with more pronounced differences seen for Tiers 2, 3 and 4. This is notable, since **+filters** removes a larger proportion of data for lower-ranked Wikipedias than those in Tier 1.

5.2 Language Adaptation

As a more common use-case for low-resource languages (Pfeiffer et al., 2020, 2022), we adapt XLM-R (Conneau et al., 2020) to Wikipedias that were not included in its training data, according to the experimental protocol of Alabi et al. (2022). Following this, we fine-tune each model on SIB200 and Belebele (Bandarkar et al., 2024) — a cross-lingual reading comprehension benchmark.[14]

The right panel of Figure 4 shows that Tier 1 Δ values remain within the $(-0.05, 0.05)$ range, mirroring the monolingual pretraining results. However, a different pattern emerges for the remaining

[12] Full training details for all 3 setups are in Appendix C.

[13] We also evaluate these models on MasakhaNER, AfriSenti-Twitter and MasakhaNEWS. However, since the results for these tasks are similar to SIB200, we include them in Appendix C in the interest of brevity.

[14] We fine-tune models on English data — as recommended in the Belebele paper — and evaluate them on the target languages in a zero shot manner. This setup is not possible with the monolingual models due to tokeniser restrictions.

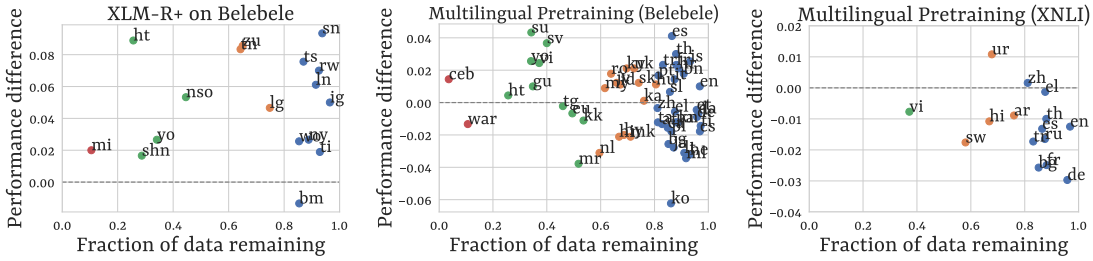

Figure 5: *Left*: Performance difference over BELEBELE after language adaptation. *Center*: Performance difference over BELEBELE after multilingual pretraining. *Right*: Performance difference over XNLI after multilingual pretraining.

tiers: most exhibit performance improvements after adapting XLM-R with **+filters** data. Acehnese is again a notable highlight here, returning $\Delta = 0.08$ — an indication that a well-calibrated model can benefit from smaller quantities of higher-quality data. We observe even better results on BELEBELE (Figure 5, left), with all but one model performing better in the **+filters** setting. Previous work has shown that quality of pretraining data is more impactful to performance in a zero-shot setting than when fine-tuning the model on the target language (Tatariya et al., 2023). These findings provide additional evidence that the filtered content was indeed low quality.

5.3 Multilingual Pretraining

In our third experiment, we mimic the setup of multilingual BERT (Devlin et al., 2019) and pretrain two base-size multilingual models (approximately 100 million parameters) on **raw_wiki** and **+filters** data for the top 100 Wikipedias sorted by raw document count. We evaluate these models on BELEBELE and XNLI (Conneau et al., 2018), after performing English task-specific fine-tuning. Regarding the former dataset, we observe similar trends to the previous two experiments: Δ values are largely within the $(-0.05, 0.05)$ range, with more lower-tier Wikipedias appearing to benefit from filtering than TIER 1 (Figure 5, center). However, we do not observe the same effect as seen for XLM-R (Figure 5, left), with all but two models dropping in performance (Figure 5, right) — albeit very slightly and within a very narrow range of $(-0.03, 0.01)$.

6 Conclusion

In this study, we demonstrated that Wikipedia data quality varies substantially across language editions, challenging the assumption that it represents a consistent and impeccably curated resource. In applying a set of common data cleaning techniques to the entire set of non-English editions, we found that approximately 30% of documents and 12% of characters thereof represent low-quality content — including varying degrees of foreign-script contamination, templates repeated across hundreds of URLs, and a proliferation of bot-generated articles. We consolidated these results into a tiered ranking of all non-English Wikipedia editions, which showed strong correspondence with alternative notions of dataset quality, such as DEPTH+, proportion of bot-generated articles, and other heuristic measures such as document length and unigram entropy. Lastly, in a downstream evaluation across three practical applications — monolingual pretraining, language adaptation, and multilingual pretraining — we confirmed that models trained on a filtered Wikipedia generally match or exceed the performance of those trained on raw data, with lower quality editions generally benefitting more.

Ultimately, the findings reported here serve to underscore the necessity of rigorous data auditing — even when working with a trusted data source like Wikipedia. While our approach is highly automated, it nonetheless demonstrates that simple, model-agnostic preprocessing can go a long way in discerning between good and bad quality data. As Wikipedia continues to serve and evolve as a foundational resource for NLP, understanding and accounting for these quality variations becomes essential for developing more equitable and effective language technologies. This is particularly true for low-resource languages, for which only several thousand (or even hundred) articles are attested. As such, we believe that future work should aim to engage with Wikipedia editing communities to better understand the sources of quality variation and identify language-specific considerations that automated filtering may overlook.

7 Limitations

Use of Encoder Models Our experimental setup has been limited to encoder-only models since already existing decoder-only models such as Llama (Touvron et al., 2023) and Deepseek (DeepSeek-AI et al., 2025) have already most likely been trained on Wikipedia. We attempted similar language adaptation experiments for decoder models, and in our experiments with Deepseek, for instance, we immediately observed low losses within the first few 100 steps of continued pretraining, indicating that the model had already been trained on Wikipedia. We also found that these models do not perform well especially for low resource languages such as the ones included in the paper (Adelani et al., 2025), with no improvement even after continued pretraining — likely due to the comparatively small size of individual Wikipedias (Zhang et al., 2024). Moreover, benchmarks for decoder models rely on English prompts to evaluate multilingual language performance, which has been noted as a problematic paradigm (see, e.g. Poelman and de Lhoneux (2025)), and which also restricts us from training and evaluating monolingual decoder-only models like Chang et al. (2024).

Evaluation Constraints Our evaluation has been restricted to benchmarks that have not been created from Wikipedia in order to not introduce additional confounds across the pretraining and fine-tuning stages. Moreover, we note that most multilingual evaluation datasets are designed to evaluate multilingual models in zero-shot or few-shot settings, without in-language train sets (like BELEBELE and XNLI), making them unusable to evaluate monolingual models due to tokenization restrictions (e.g. a tokenizer trained on English text cannot transfer to Japanese by design, as both languages employ different alphabets). Our evaluation setup is, as a consequence, limited.

8 Acknowledgements

For KT, AK, WP and MDL the computational resources and services used were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI. AK was partially funded by a Google Research Award. WP is funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096. EP, HL and JB are funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate programme (project nr. CF210454). MB is supported by TrustLLM funded by Horizon Europe GA 101135671, with computational resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) partially funded by the Swedish Research Council through grant agreement no. 2024/22-745. We would also like to thank Isaac Caswell, Colin Cherry, and Markus Freitag at Google Translate Research for their insightful reviews on the initial draft.

References

01.AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, and 12 others. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of 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 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings*

of the 29th International Conference on Computational Linguistics, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Saied Alshahrani, Norah Alshahrani, and Jeanna Matthews. 2023. DEPTH+: An enhanced depth metric for Wikipedia corpora quality. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 175–189, Toronto, Canada. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *Preprint*, arXiv:2408.10441.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in NLP. *Preprint*, arXiv:2311.08391.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany. Association for Computational Linguistics.

Peter Izsak, Moshe Berchansky, and Omer Levy. 2021. How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jacob Judah. 2025. How AI and Wikipedia have sent vulnerable languages into a doom spiral.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

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

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *ArXiv*, abs/1901.07291.

Jens Lehmann, Dhananjay Bhandiwad, Preetam Gattogi, and Sahar Vahdati. 2024. Beyond boundaries: A human-like approach for question answering over structured and unstructured information sources. *Transactions of the Association for Computational Linguistics*, 12:786–802.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, and 2 others. 2024. Creoleval: Multilingual multitask benchmarks for creoles. *Preprint*, arXiv:2310.19567.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 523–532, Dublin, Ireland. Association for Computational Linguistics.

Ivan Lokhov. 2021. Why are there so many Wikipedia articles in Swedish and Cebuano? - Datawrapper Blog.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis. *Preprint*, arXiv:2306.07377.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 7654–7673, Online. Association for Computational Linguistics.

Wessel Poelman and Miryam de Lhoneux. 2025. The roles of English in evaluating multilingual language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 492–498, Tallinn, Estonia. University of Tartu Library.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Kushal Tatariya, Heather Lent, and Miryam de Lhoneux. 2023. Transfer learning for code-mixed data: Do pretraining languages matter? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 365–378, Toronto, Canada. Association for Computational Linguistics.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *Preprint*, arXiv:2308.12284.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Rik van Noord, Taja Kuzman, Peter Rupnik, Nikola Ljubešić, Miquel Esplà-Gomis, Gema Ramírez-Sánchez, and Antonio Toral. 2024. Do language models care about text quality? evaluating web-crawled corpora across 11 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5221–5234, Torino, Italia. ELRA and ICCL.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

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.

Kyle Wilson. 2019. Wikipedia has a Google Translate problem.

Mike Zhang, Max Müller-Eberstein, Elisa Bassignana, and Rob van der Goot. 2024. Snakmodel: Lessons learned from training an open danish large language model. *Preprint*, arXiv:2412.12956.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, Massachusetts.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

A Script Filtering

For script filtering, we utilize a regular expression that matches unicode characters belonging to the Wikipedia-documented ISO-15924 script range (e.g. `Latn`), as well as a set of neutral categories (e.g. Punctuation, Symbols, Numbers, etc.). After filtering out non-official scripts, we run a post-hoc clean-up filter that removes trailing whitespace, empty bracket groups, etc. Under this setup, foreign-script named entities and references are removed, and ones in the official script(s) are retained.

B Language-specific Heuristic Thresholding

Filtering Wikipedia using heuristics requires an arbitrary threshold to serve as a cut-off between 'good' and 'bad' quality. Existing thresholds (e.g. Gopher (Rae et al., 2021)) are normally tuned for English, and not directly applicable to other languages. Though we attempted to filter Wikipedia articles on thresholds defined by language-specific data distributions and conducted similar downstream experiments as detailed in §5, we found little gains over simple script filtering and deduplication. We detail the procedure for creating language-specific thresholds for heuristic metrics below:

For a given Wikipedia metric distribution (e.g., `rps_doc_word_count`), we determine a representative sample size based on the full dataset. Testing across various metrics and Wikipedias, we select a sample size of $n_{\text{sample}} = 0.05 * n_{\text{docs}}$, where n_{docs} is the total number of documents in the dataset. We then sort the full distribution in ascending order and take the first n_{sample} values as the *value distribution*, D_{value}. This distribution reflects the lowest values in the distribution —essentially, the values we want to assess as potential outliers or benchmarks. Separately, we randomly sample n_{sample} values from the full distribution to create the *sampled distribution*, D_{sample}. This random sample provides a baseline density that represents the overall data distribution rather than just extreme values. We estimate the kernel density (KDE) for both D_{value} and D_{sample}, applying these KDEs over n_{sample} evenly spaced values in the range:

$$\left[\min(D_{\text{value}}),\ \max(D_{\text{sample}})\right]$$

The threshold is then identified as the point within this range where the difference between the two density estimates is maximized.

C Training Hyperparameters and Finetuning Results

Hyperparameters for the pretraining and finetuning of monolingual models are in Table 2. We use the SentencePiece package (Kudo and Richardson, 2018) to train monolingual unigram tokenisers for each language. We train two tokenisers per language: one on the **raw_wiki** with no filtering, which is used to train the **raw_wiki** and **random** models; the second one on the data remaining after script filtering and deduplication, which we use to train the **+filters** models. We use an approximation with Zipf's law (Zipf, 1949) on each Wikipedia to estimate the vocabulary size for each model, mentioned in Table 5. Hyperparameters for XLM-R language adaptation are in Table 3. For both cases, we checkpoint and use the best model on a held-out validation set to fine-tune after pretraining. For multilingual pretraining, we train 2 WordPiece tokenizers on **raw_wiki** and **+filters** with a vocabulary size of 110k. Languages are listed in Table 6 and hyperparameters in Table 4.

Tiny DeBERTa Hyperparameters	
Pretraining	
Model Type	deberta
Hidden Size	312
Intermediate Size	1200
Attention Heads	12
Hidden Layers	4
Hidden Act	gelu
Max Length	128
Task	MLM
Mask Probability	0.4
Padding Strategy	longest
Grad. Accumulation Steps	4
Batch size	8 to 32
Learning Rate	1.00e-3
Epochs	100
Finetuning	
Max Length	128
Epochs	20
Batch Size	8
Learning Rate	5.00e-5
Padding Strategy	max_length

Table 2: Hyperparameters for monolingual pretraining and fine-tuning. While we used 100 epochs as a default, for Wikipedias that are too large (`nl, ro, id, hu, uk, vi, ceb, war, tt, azb, min`) we pretrain only for 2 epochs. Similarly, while we used a batch size of 8 for all models, for the larger Wikipedias we use a batch size of 32. All models are trained on 1 GPU.

XLMR Hyperparameters	
Language Adaptation	
Model Type	xlm-roberta
Task	MLM
Max Length	512
Mask Probability	0.15
Epochs	3
Batch Size	8
Learning Rate	5.00e-5
Padding strategy	Longest
Grad. Accumulation Steps	1
Finetuning - SIB200	
Max Length	512
Epochs	20
Batch Size	32
Learning Rate	5.00e-5
Padding Strategy	max_length
Finetuning - Belebele	
Max Length	128
Epochs	3
Batch Size	8
Learning Rate	5.00e-5
Padding Strategy	max_length

Table 3: Hyperparameters for XLM-R language adaptation (based on Alabi et al. (2022)) and fine-tuning.

Multilingual Pretraining	
Pretraining	
Sampling Temperature	1.43
Model Type	bert
Task	MLM
Max Length	128
Mask Probability	0.12
Steps	200,000
Batch Size (per device)	256
Learning Rate	1.00e-3
Padding strategy	Longest
Grad. Accumulation Steps	4
Finetuning - Belebele	
Max Length	128
Epochs	3
Batch Size	8
Learning Rate	5.00e-5
Padding Strategy	max_length
Finetuning - MNLI	
Max Length	128
Epochs	3
Batch Size	32
Learning Rate	2.00e-5
Padding Strategy	max_length

Table 4: Hyperparameters for multilingual pretraining and fine-tuning, based on mBERT and Izsak et al. (2021). Both models are trained on 4 H100 GPUs.

Estimated Vocabulary Size			
Lang	**Vocab**	**Wiki-ID**	**Vocab**
ace	3505	mr	14708
am	5234	nl	60011
ary	4502	nso	2801
awa	2190	ny	2472
azb	13996	om	3436
ban	6624	pag	2170
bm	1485	pcm	2595
ceb	73124	rn	1446
ckb	11451	ro	36329
ee	1770	rw	5725
fon	1426	shn	5784
ha	13126	sn	5227
ht	10836	so	6117
hu	46205	sw	12529
id	41725	te	23175
ig	11609	ti	1150
kk	22496	tn	3783
ks	2248	ts	1817
lg	4528	tt	26047
lij	5462	tw	4745
ln	2509	uk	62626
mai	4793	vi	43806
mi	3451	war	25803
min	14783	wo	3384
mni	3351	yo	6128
		zu	4439

Table 5: Estimated vocabulary size used to train monolingual tokenisers for each Wikipedia in our experimental pipeline.

Languages Included in Multilingual Pretraining			
Language	**Wiki-ID**	**Language**	**Wiki-ID**
Afrikaans	af	Korean	ko
Albanian	sq	Latin	la
Arabic	ar	Latvian	lv
Aragonese	an	Lithuanian	lt
Armenian	hy	Lombard	lmo
Asturian	ast	Low German	nds-nl
Azerbaijani	az	Luxembourgish	lb
Bashkir	ba	Macedonian	mk
Basque	eu	Malagasy	mg
Bavarian	bar	Malay	ms
Belarusian	be	Malayalam	ml
Bangla	bn	Marathi	mr
Bishnupriya	bpy	Minangkabau	min
Bosnian	bs	Nepali	ne
Breton	br	Newari	new
Bulgarian	bg	Norwegian Nynorsk	nn
Burmese	my	Occitan	oc
Catalan	ca	Persian	fa
Chechen	ce	Piedmontese	pms
Chinese	zh	Polish	pl
Cantonese	zh-yue	Portuguese	pt
Cebuano	ceb	Punjabi	pa
Chuvash	cv	Romanian	ro
Croatian	hr	Rusyn	rue
Czech	cs	Scots	sco
Danish	da	Serbo-Croatian	sh
Dutch	nl	Sicilian	scn
English	en	Slovak	sk
Estonian	et	Slovenian	sl
Finnish	fi	South Azerbaijani	azb
French	fr	Spanish	es
Galician	gl	Sundanese	su
Georgian	ka	Swahili	sw
German	de	Swedish	sv
Greek	el	Filipino	tl
Gujarati	gu	Tajik	tg
Haitian Creole	ht	Tamil	ta
Hebrew	he	Tatar	tt
Hindi	hi	Turkish	tr
Hungarian	hu	Ukrainian	uk
Icelandic	is	Uzbek	uz
Ido	io	Vietnamese	vi
Indonesian	id	Volapük	vo
Irish	ga	Waray	war
Italian	it	Welsh	cy
Japanese	ja	Western Frisian	fy
Javanese	jv	Western Panjabi	pnb
Kannada	kn	Yoruba	yo
Kazakh	kk	Thai	th
Kyrgyz	ky	Mongolian	mn

Table 6: Wikipedias included in the pretraining of the multilingual models.

	SIB200		
wiki	**raw_wiki**	**+filters**	**random**
	TIER 1		
ti	$0.38_{\pm 0.05}$	$0.38_{\pm 0.01}$	$0.32_{\pm 0.01}$
ha	$0.77_{\pm 0.02}$	$\mathbf{0.78}_{\pm 0.02}$	$0.62_{\pm 0.03}$
ts	$\mathbf{0.61}_{\pm 0.03}$	$0.60_{\pm 0.01}$	$0.53_{\pm 0.02}$
ig	$0.76_{\pm 0.01}$	$\mathbf{0.77}_{\pm 0.02}$	$0.64_{\pm 0.02}$
sn	$\mathbf{0.66}_{\pm 0.03}$	$0.62_{\pm 0.03}$	$0.43_{\pm 0.05}$
so	$0.68_{\pm 0.01}$	$\mathbf{0.73}_{\pm 0.02}$	$0.56_{\pm 0.06}$
rw	$0.77_{\pm 0.01}$	$\mathbf{0.77}_{\pm 0.01}$	$0.65_{\pm 0.03}$
tw	$\mathbf{0.65}_{\pm 0.10}$	$0.65_{\pm 0.02}$	$0.63_{\pm 0.03}$
ln	$\mathbf{0.59}_{\pm 0.02}$	$0.57_{\pm 0.02}$	$0.52_{\pm 0.04}$
wo	$0.50_{\pm 0.02}$	$\mathbf{0.52}_{\pm 0.02}$	$0.45_{\pm 0.04}$
om	$0.59_{\pm 0.01}$	$\mathbf{0.62}_{\pm 0.01}$	$0.34_{\pm 0.05}$
ny	$\mathbf{0.64}_{\pm 0.02}$	$0.60_{\pm 0.01}$	$0.60_{\pm 0.01}$
ee	$0.59_{\pm 0.03}$	$\mathbf{0.61}_{\pm 0.03}$	$0.57_{\pm 0.01}$
bm	$0.45_{\pm 0.02}$	$\mathbf{0.48}_{\pm 0.02}$	$0.47_{\pm 0.03}$
fon	$0.48_{\pm 0.03}$	$0.45_{\pm 0.04}$	$0.44_{\pm 0.01}$
Avg	0.61	**0.61**	0.51
	TIER 2		
ro	$\mathbf{0.82}_{\pm 0.03}$	$0.80_{\pm 0.02}$	$0.66_{\pm 0.01}$
id	$0.83_{\pm 0.01}$	$\mathbf{0.85}_{\pm 0.01}$	$0.75_{\pm 0.04}$
hu	$0.82_{\pm 0.01}$	$\mathbf{0.86}_{\pm 0.01}$	$0.64_{\pm 0.05}$
nl	$\mathbf{0.82}_{\pm 0.03}$	$0.80_{\pm 0.03}$	$0.66_{\pm 0.03}$
uk	$\mathbf{0.85}_{\pm 0.01}$	$0.81_{\pm 0.02}$	$0.43_{\pm 0.08}$
awa	$0.54_{\pm 0.03}$	$0.52_{\pm 0.03}$	$0.47_{\pm 0.02}$
ks	$0.42_{\pm 0.02}$	$\mathbf{0.45}_{\pm 0.02}$	$0.41_{\pm 0.05}$
ckb	$0.80_{\pm 0.01}$	$\mathbf{0.82}_{\pm 0.01}$	$0.54_{\pm 0.04}$
rn	$0.51_{\pm 0.03}$	$0.50_{\pm 0.02}$	$0.45_{\pm 0.03}$
sw	$0.79_{\pm 0.02}$	$0.81_{\pm 0.02}$	$0.58_{\pm 0.05}$
am	$\mathbf{0.66}_{\pm 0.01}$	$0.65_{\pm 0.02}$	$0.41_{\pm 0.02}$
zu	$\mathbf{0.73}_{\pm 0.01}$	$0.70_{\pm 0.02}$	$0.61_{\pm 0.02}$
tn	$0.64_{\pm 0.02}$	$0.61_{\pm 0.01}$	$0.37_{\pm 0.05}$
ary	$0.59_{\pm 0.00}$	$0.60_{\pm 0.01}$	$0.46_{\pm 0.02}$
lg	$0.55_{\pm 0.02}$	$0.55_{\pm 0.02}$	$0.51_{\pm 0.02}$
Avg	0.69	0.69	0.53
	TIER 3		
lij	$\mathbf{0.61}_{\pm 0.01}$	$0.59_{\pm 0.02}$	$0.45_{\pm 0.02}$
ht	$\mathbf{0.70}_{\pm 0.04}$	$0.69_{\pm 0.01}$	$0.36_{\pm 0.03}$
te	$0.80_{\pm 0.02}$	$0.79_{\pm 0.02}$	$0.52_{\pm 0.03}$
mr	$0.74_{\pm 0.04}$	$\mathbf{0.78}_{\pm 0.02}$	$0.50_{\pm 0.04}$
kk	$0.81_{\pm 0.02}$	$\mathbf{0.83}_{\pm 0.02}$	$0.74_{\pm 0.02}$
yo	$\mathbf{0.70}_{\pm 0.02}$	$0.69_{\pm 0.02}$	$0.52_{\pm 0.07}$
ace	$\mathbf{0.27}_{\pm 0.02}$	$0.15_{\pm 0.03}$	$0.23_{\pm 0.01}$
shn	$0.74_{\pm 0.02}$	$0.72_{\pm 0.02}$	$0.62_{\pm 0.04}$
mai	$0.67_{\pm 0.01}$	$0.68_{\pm 0.02}$	$0.55_{\pm 0.02}$
ban	$0.65_{\pm 0.01}$	$0.66_{\pm 0.02}$	$0.49_{\pm 0.03}$
vi	$0.82_{\pm 0.01}$	$\mathbf{0.83}_{\pm 0.00}$	$0.74_{\pm 0.02}$
nso	$0.50_{\pm 0.02}$	$\mathbf{0.53}_{\pm 0.01}$	$0.46_{\pm 0.04}$
Avg	**0.67**	0.66	0.52
	TIER 4		
tt	$\mathbf{0.79}_{\pm 0.01}$	$0.75_{\pm 0.01}$	$0.48_{\pm 0.03}$
ceb	$0.39_{\pm 0.03}$	$\mathbf{0.51}_{\pm 0.03}$	$0.28_{\pm 0.04}$
war	$0.48_{\pm 0.04}$	$0.49_{\pm 0.04}$	$0.46_{\pm 0.02}$
pag	$0.45_{\pm 0.03}$	$\mathbf{0.48}_{\pm 0.01}$	$0.35_{\pm 0.02}$
mi	$\mathbf{0.55}_{\pm 0.02}$	$0.52_{\pm 0.02}$	$0.48_{\pm 0.03}$
mni	$\mathbf{0.27}_{\pm 0.04}$	$0.21_{\pm 0.01}$	$0.31_{\pm 0.00}$
azb	$\mathbf{0.53}_{\pm 0.05}$	$0.53_{\pm 0.02}$	$0.42_{\pm 0.03}$
min	$\mathbf{0.65}_{\pm 0.03}$	$0.60_{\pm 0.03}$	$0.41_{\pm 0.02}$
Avg	**0.51**	0.51	0.40

Table 7: Fine-tuning results on SIB200 for tiny DeBERTa models. Results are averaged over 5 runs.

MasakhaNER 2			
Lang	**raw_wiki**	**+filters**	**random**
pcm	**0.77**$_{\pm 0.00}$	0.76$_{\pm 0.00}$	0.74$_{\pm 0.00}$
ha	**0.82**$_{\pm 0.00}$	0.81$_{\pm 0.00}$	0.75$_{\pm 0.01}$
ig	0.85$_{\pm 0.00}$	**0.85**$_{\pm 0.00}$	0.81$_{\pm 0.01}$
sn	**0.92**$_{\pm 0.00}$	0.92$_{\pm 0.00}$	0.90$_{\pm 0.00}$
rw	0.76$_{\pm 0.05}$	**0.79**$_{\pm 0.01}$	0.74$_{\pm 0.01}$
tw	0.76$_{\pm 0.00}$	**0.76**$_{\pm 0.01}$	0.71$_{\pm 0.00}$
wo	**0.75**$_{\pm 0.01}$	0.73$_{\pm 0.01}$	0.73$_{\pm 0.01}$
ny	0.87$_{\pm 0.00}$	**0.87**$_{\pm 0.00}$	0.85$_{\pm 0.00}$
ee	0.81$_{\pm 0.00}$	**0.81**$_{\pm 0.01}$	0.78$_{\pm 0.01}$
bm	**0.74**$_{\pm 0.01}$	0.74$_{\pm 0.01}$	0.73$_{\pm 0.00}$
fon	0.74$_{\pm 0.01}$	0.74$_{\pm 0.00}$	0.71$_{\pm 0.01}$
sw	0.90$_{\pm 0.00}$	0.90$_{\pm 0.00}$	0.88$_{\pm 0.00}$
zu	**0.83**$_{\pm 0.01}$	0.82$_{\pm 0.01}$	0.77$_{\pm 0.01}$
lg	0.84$_{\pm 0.00}$	**0.85**$_{\pm 0.00}$	0.80$_{\pm 0.01}$
tn	**0.80**$_{\pm 0.01}$	0.80$_{\pm 0.00}$	0.63$_{\pm 0.01}$
yo	0.84$_{\pm 0.00}$	**0.84**$_{\pm 0.00}$	0.80$_{\pm 0.00}$
Avg	0.81	0.81	0.77

Table 8: Fine-tuning results on MasakhaNER 2. Results are averaged over 5 runs, and languages are grouped by tier.

MasakhaNEWS			
wiki	**raw_wiki**	**+filters**	**random**
pcm	0.90$_{\pm 0.04}$	**0.91**$_{\pm 0.00}$	0.88$_{\pm 0.02}$
ha	**0.88**$_{\pm 0.01}$	0.86$_{\pm 0.01}$	0.84$_{\pm 0.01}$
ig	**0.85**$_{\pm 0.01}$	0.85$_{\pm 0.01}$	0.81$_{\pm 0.01}$
sn	0.89$_{\pm 0.01}$	**0.90**$_{\pm 0.01}$	0.87$_{\pm 0.01}$
so	0.71$_{\pm 0.01}$	**0.71**$_{\pm 0.01}$	0.59$_{\pm 0.03}$
ln	**0.78**$_{\pm 0.01}$	0.76$_{\pm 0.01}$	0.74$_{\pm 0.02}$
om	0.67$_{\pm 0.30}$	**0.82**$_{\pm 0.01}$	0.71$_{\pm 0.04}$
ti	0.60$_{\pm 0.04}$	**0.65**$_{\pm 0.02}$	0.56$_{\pm 0.02}$
sw	0.81$_{\pm 0.01}$	0.82$_{\pm 0.00}$	0.75$_{\pm 0.01}$
am	0.84$_{\pm 0.04}$	**0.87**$_{\pm 0.01}$	0.83$_{\pm 0.01}$
lg	0.86$_{\pm 0.01}$	**0.86**$_{\pm 0.02}$	0.84$_{\pm 0.02}$
rn	**0.75**$_{\pm 0.01}$	0.76$_{\pm 0.02}$	0.71$_{\pm 0.01}$
yo	0.84$_{\pm 0.03}$	0.86$_{\pm 0.01}$	0.82$_{\pm 0.01}$
Avg	0.8	**0.82**	0.76

Table 9: Fine-tuning results on MasakhaNEWS. Results are averaged over 5 runs, and languages are grouped by tier.

AfriSenti-Twitter			
wiki	**raw_wiki**	**+filters**	**random**
pcm	0.61$_{\pm 0.01}$	**0.61**$_{\pm 0.01}$	0.60$_{\pm 0.01}$
ha	0.72$_{\pm 0.01}$	0.72$_{\pm 0.01}$	0.72$_{\pm 0.01}$
ig	0.74$_{\pm 0.00}$	0.75$_{\pm 0.00}$	0.74$_{\pm 0.01}$
rw	0.60$_{\pm 0.04}$	**0.61**$_{\pm 0.01}$	0.60$_{\pm 0.01}$
tw	0.62$_{\pm 0.01}$	**0.64**$_{\pm 0.01}$	0.63$_{\pm 0.01}$
ts	0.48$_{\pm 0.04}$	0.49$_{\pm 0.03}$	**0.50**$_{\pm 0.01}$
sw	0.55$_{\pm 0.03}$	**0.59**$_{\pm 0.02}$	0.53$_{\pm 0.04}$
am	0.40$_{\pm 0.07}$	0.41$_{\pm 0.04}$	0.45$_{\pm 0.10}$
ary	**0.46**$_{\pm 0.01}$	0.40$_{\pm 0.01}$	0.44$_{\pm 0.01}$
yo	0.68$_{\pm 0.01}$	**0.68**$_{\pm 0.01}$	0.64$_{\pm 0.01}$
Avg	0.59	**0.59**	0.59

Table 10: Fine-tuning results on AfriSenti-Twitter. Results are averaged over 5 runs, and languages are grouped by tier.

XLMR Continued Pretraining (SIB200)			
wiki	**baseline**	**raw_wiki**	**+filters**
ti	0.59$_{\pm 0.04}$	0.53$_{\pm 0.21}$	0.56$_{\pm 0.02}$
fon	0.43$_{\pm 0.20}$	**0.60**$_{\pm 0.04}$	0.56$_{\pm 0.02}$
ig	0.52$_{\pm 0.20}$	**0.79**$_{\pm 0.02}$	**0.79**$_{\pm 0.03}$
ee	0.64$_{\pm 0.04}$	**0.66**$_{\pm 0.03}$	0.64$_{\pm 0.03}$
ny	0.64$_{\pm 0.01}$	**0.68**$_{\pm 0.03}$	0.65$_{\pm 0.02}$
wo	0.60$_{\pm 0.03}$	**0.65**$_{\pm 0.01}$	0.64$_{\pm 0.01}$
ln	0.66$_{\pm 0.01}$	**0.70**$_{\pm 0.01}$	0.69$_{\pm 0.01}$
tw	0.66$_{\pm 0.02}$	**0.72**$_{\pm 0.02}$	0.70$_{\pm 0.03}$
rw	0.56$_{\pm 0.03}$	**0.69**$_{\pm 0.02}$	0.67$_{\pm 0.02}$
sn	0.54$_{\pm 0.03}$	0.59$_{\pm 0.02}$	**0.60**$_{\pm 0.03}$
bm	0.43$_{\pm 0.22}$	**0.60**$_{\pm 0.05}$	0.55$_{\pm 0.03}$
ts	0.58$_{\pm 0.05}$	**0.63**$_{\pm 0.04}$	0.60$_{\pm 0.05}$
Avg	0.57	0.65	0.64
tn	0.56$_{\pm 0.02}$	**0.62**$_{\pm 0.03}$	0.60$_{\pm 0.04}$
ks	0.62$_{\pm 0.22}$	**0.68**$_{\pm 0.02}$	**0.68**$_{\pm 0.02}$
rn	0.54$_{\pm 0.04}$	0.55$_{\pm 0.06}$	**0.58**$_{\pm 0.02}$
lg	0.40$_{\pm 0.17}$	**0.58**$_{\pm 0.03}$	**0.58**$_{\pm 0.03}$
zu	0.58$_{\pm 0.01}$	0.64$_{\pm 0.03}$	**0.68**$_{\pm 0.02}$
awa	0.80$_{\pm 0.04}$	0.81$_{\pm 0.02}$	**0.83**$_{\pm 0.03}$
Avg	0.59	0.65	0.66
ht	0.63$_{\pm 0.03}$	**0.73**$_{\pm 0.03}$	**0.73**$_{\pm 0.02}$
nso	0.55$_{\pm 0.02}$	0.59$_{\pm 0.04}$	**0.63**$_{\pm 0.02}$
lij	0.77$_{\pm 0.02}$	**0.81**$_{\pm 0.03}$	0.78$_{\pm 0.03}$
ace	0.41$_{\pm 0.09}$	0.45$_{\pm 0.13}$	**0.53**$_{\pm 0.06}$
shn	0.33$_{\pm 0.16}$	0.61$_{\pm 0.03}$	0.60$_{\pm 0.04}$
mai	0.82$_{\pm 0.02}$	0.83$_{\pm 0.02}$	0.82$_{\pm 0.02}$
ban	0.79$_{\pm 0.01}$	0.82$_{\pm 0.02}$	**0.83**$_{\pm 0.01}$
yo	0.56$_{\pm 0.02}$	0.63$_{\pm 0.04}$	**0.66**$_{\pm 0.02}$
Avg	0.61	0.68	0.70
mi	0.59$_{\pm 0.12}$	0.62$_{\pm 0.02}$	**0.63**$_{\pm 0.03}$
mni	**0.54**$_{\pm 0.05}$	0.38$_{\pm 0.11}$	0.45$_{\pm 0.03}$
pag	0.77$_{\pm 0.02}$	**0.79**$_{\pm 0.02}$	0.77$_{\pm 0.05}$
Avg	0.59	0.66	0.66

Table 11: XLM-R base and XLM-R with continued pretraining on unseen languages with **raw_wiki** and **+filters** fine-tuned on SIB200. Results are grouped by tier.

XLMR Continued Pretraining (Belebele)			
wiki	**baseline**	**raw_wiki**	**+filters**
wo	0.27	0.26	**0.29**
ts	0.24	0.25	**0.33**
ny	0.26	0.25	**0.27**
rw	0.24	0.23	**0.30**
ig	0.26	0.25	**0.30**
ln	0.25	0.24	**0.30**
sn	0.25	0.23	**0.33**
bm	0.24	**0.26**	0.24
ti	0.24	0.27	**0.29**
lg	0.24	0.24	**0.28**
tn	0.25	0.24	**0.33**
zu	0.25	0.23	**0.32**
shn	0.25	0.26	**0.27**
nso	0.25	0.26	**0.31**
ht	0.24	0.24	**0.33**
yo	0.26	0.25	**0.27**
mi	0.26	0.27	**0.29**
Avg	0.25	0.25	**0.30**

Table 12: XLM-R base and XLM-R with continued pretraining on unseen languages with **raw_wiki** and **+filters** fine-tuned on MC and tested on BELEBELE. Languages are grouped by tier.

Multilingual Pretraining (XNLI)		
wiki	**raw_wiki**	**+filters**
el	0.39	0.39
bg	0.42	0.40
de	0.46	0.43
tr	0.43	0.41
th	0.38	0.37
fr	0.48	0.46
ru	0.40	0.38
es	0.50	0.49
zh	0.43	0.43
en	0.72	0.71
ar	0.39	0.38
sw	0.41	0.39
ur	0.38	0.39
hi	0.38	0.37
vi	0.41	0.41
Avg	0.44	0.43

Table 13: Multilingual models pretrained on top 100 Wikipedias with **raw_wiki** and **+filters** fine-tuned on MNLI and tested on XNLI. Languages are grouped by tier.

Multilingual Pretraining (Belebele)		
wiki	**raw_wiki**	**+filters**
pa	0.27	0.26
ta	0.24	0.23
af	0.24	0.23
th	0.22	0.25
fr	0.24	0.26
es	0.22	0.26
cs	0.26	0.24
pl	0.26	0.24
de	0.27	0.26
en	0.25	0.26
fi	0.26	0.25
pt	0.25	0.27
hr	0.25	0.27
el	0.24	0.23
da	0.26	0.25
it	0.25	0.26
bn	0.25	0.27
ml	0.27	0.24
is	0.25	0.27
bg	0.27	0.24
kn	0.26	0.25
et	0.27	0.26
ca	0.28	0.26
ko	0.27	0.20
tr	0.24	0.26
he	0.27	0.24
zh	0.25	0.24
sl	0.25	0.25
ja	0.27	0.24
lt	0.26	0.23
sk	0.23	0.24
ky	0.24	0.26
ka	0.25	0.25
mk	0.27	0.25
uk	0.24	0.26
hu	0.25	0.27
ro	0.25	0.27
id	0.26	0.27
hi	0.26	0.24
jv	0.23	0.24
hy	0.26	0.24
my	0.23	0.24
nl	0.26	0.23
sv	0.22	0.26
yo	0.23	0.26
vi	0.24	0.26
tg	0.24	0.24
su	0.24	0.28
gu	0.23	0.24
eu	0.26	0.25
ht	0.26	0.26
kk	0.27	0.26
mr	0.28	0.24
ceb	0.25	0.26
war	0.24	0.23
Avg	0.25	0.25

Table 14: Multilingual models pretrained on top 100 Wikipedias with **raw_wiki** and **+filters** fine-tuned on MC and tested on BELEBELE. Languages are grouped by tier.

D Quality Tiers

Wiki-Id	Tier	Wiki-Id	Tier	Wiki-Id	Tier	Wiki-Id	Tier	Wiki-Id	Tier	Wiki-Id	Tier
ab	Tier 3	crh	Tier 3	ha	Tier 1	lez	Tier 1	pa	Tier 1	su	Tier 3
ace	Tier 3	cs	Tier 1	hak	Tier 3	lfn	Tier 1	pag	Tier 4	sv	Tier 3
ady	Tier 1	csb	Tier 1	haw	Tier 3	lg	Tier 2	pam	Tier 2	sw	Tier 2
af	Tier 1	cu	Tier 1	he	Tier 1	li	Tier 1	pap	Tier 1	szl	Tier 3
als	Tier 1	cv	Tier 4	hi	Tier 2	lij	Tier 3	pcd	Tier 1	ta	Tier 1
alt	Tier 2	cy	Tier 2	hif	Tier 2	lld	Tier 4	pcm	Tier 1	tay	Tier 2
am	Tier 2	da	Tier 1	hr	Tier 1	lmo	Tier 3	pdc	Tier 1	tcy	Tier 1
ami	Tier 2	dag	Tier 3	hsb	Tier 2	ln	Tier 1	pfl	Tier 3	te	Tier 3
an	Tier 1	de	Tier 1	ht	Tier 3	lo	Tier 2	pi	Tier 4	tet	Tier 1
ang	Tier 1	din	Tier 1	hu	Tier 2	lt	Tier 1	pih	Tier 1	tg	Tier 3
ar	Tier 2	diq	Tier 3	hy	Tier 2	ltg	Tier 1	pl	Tier 1	th	Tier 1
arc	Tier 1	dsb	Tier 1	hyw	Tier 1	lv	Tier 1	pms	Tier 3	ti	Tier 1
arz	Tier 3	dty	Tier 2	ia	Tier 3	mad	Tier 1	pnb	Tier 2	tk	Tier 2
as	Tier 1	dv	Tier 1	id	Tier 2	mai	Tier 3	pnt	Tier 2	tl	Tier 1
ast	Tier 2	dz	Tier 1	ie	Tier 3	map-bms	Tier 3	ps	Tier 1	tly	Tier 3
atj	Tier 2	ee	Tier 1	ig	Tier 1	mdf	Tier 3	pt	Tier 1	tn	Tier 2
av	Tier 2	el	Tier 1	ik	Tier 1	mg	Tier 3	pwn	Tier 1	to	Tier 1
awa	Tier 2	eml	Tier 3	ilo	Tier 2	mhr	Tier 2	qu	Tier 2	tpi	Tier 2
ay	Tier 1	eo	Tier 1	inh	Tier 1	mi	Tier 4	rm	Tier 2	tr	Tier 1
az	Tier 2	es	Tier 1	io	Tier 2	min	Tier 4	rmy	Tier 1	trv	Tier 1
azb	Tier 4	et	Tier 1	is	Tier 1	mk	Tier 2	rn	Tier 2	ts	Tier 1
ba	Tier 2	eu	Tier 3	it	Tier 1	ml	Tier 1	ro	Tier 2	tt	Tier 4
ban	Tier 3	ext	Tier 1	iu	Tier 2	mn	Tier 1	roa-rup	Tier 1	tum	Tier 3
bar	Tier 1	fa	Tier 2	ja	Tier 1	mni	Tier 4	roa-tara	Tier 2	tw	Tier 1
bat-smg	Tier 3	fat	Tier 1	jam	Tier 1	mnw	Tier 1	ru	Tier 1	ty	Tier 3
bcl	Tier 2	ff	Tier 1	jbo	Tier 1	mr	Tier 3	rue	Tier 2	tyv	Tier 1
be	Tier 2	fi	Tier 1	jv	Tier 2	mrj	Tier 2	rw	Tier 1	udm	Tier 2
be-x-old	Tier 1	fiu-vro	Tier 1	ka	Tier 2	ms	Tier 3	sa	Tier 2	ug	Tier 2
bg	Tier 1	fj	Tier 1	kaa	Tier 1	mt	Tier 1	sah	Tier 1	uk	Tier 2
bh	Tier 2	fo	Tier 1	kab	Tier 1	mwl	Tier 1	sat	Tier 1	ur	Tier 2
bi	Tier 1	fon	Tier 1	kbd	Tier 1	my	Tier 2	sc	Tier 1	uz	Tier 2
bjn	Tier 3	fr	Tier 1	kbp	Tier 1	myv	Tier 2	scn	Tier 2	ve	Tier 3
blk	Tier 1	frp	Tier 1	kcg	Tier 1	mzn	Tier 3	sco	Tier 1	vec	Tier 3
bm	Tier 1	frr	Tier 1	kg	Tier 2	nah	Tier 2	sd	Tier 1	vep	Tier 1
bn	Tier 1	fur	Tier 1	ki	Tier 2	nap	Tier 3	se	Tier 2	vi	Tier 3
bo	Tier 2	fy	Tier 1	kk	Tier 3	nds	Tier 1	sg	Tier 1	vls	Tier 1
bpy	Tier 3	ga	Tier 1	kl	Tier 1	ne	Tier 1	sh	Tier 3	vo	Tier 2
br	Tier 1	gag	Tier 2	km	Tier 1	new	Tier 4	shi	Tier 2	wa	Tier 1
bs	Tier 2	gan	Tier 3	kn	Tier 1	nia	Tier 2	shn	Tier 3	war	Tier 4
bug	Tier 4	gcr	Tier 1	ko	Tier 1	nl	Tier 2	si	Tier 1	wo	Tier 1
bxr	Tier 1	gd	Tier 1	koi	Tier 3	nn	Tier 1	simple	Tier 1	wuu	Tier 1
ca	Tier 1	gl	Tier 1	krc	Tier 2	no	Tier 1	sk	Tier 2	xal	Tier 3
cbk-zam	Tier 3	glk	Tier 3	ks	Tier 2	nov	Tier 2	sl	Tier 1	xh	Tier 2
cdo	Tier 3	gn	Tier 1	ksh	Tier 1	nrm	Tier 2	sm	Tier 1	xmf	Tier 2
ce	Tier 4	gom	Tier 1	ku	Tier 3	nso	Tier 3	sn	Tier 1	yi	Tier 1
ceb	Tier 4	gor	Tier 3	kv	Tier 2	nv	Tier 3	so	Tier 1	yo	Tier 3
ch	Tier 2	got	Tier 2	kw	Tier 1	ny	Tier 1	sq	Tier 2	za	Tier 2
chr	Tier 3	gu	Tier 3	ky	Tier 2	oc	Tier 2	sr	Tier 3	zea	Tier 2
chy	Tier 1	guc	Tier 1	la	Tier 2	olo	Tier 1	srn	Tier 3	zh	Tier 1
ckb	Tier 2	gur	Tier 1	lad	Tier 1	om	Tier 1	ss	Tier 1	zh-classical	Tier 1
co	Tier 1	guw	Tier 1	lb	Tier 1	or	Tier 2	st	Tier 2	zh-min-nan	Tier 4
cr	Tier 3	gv	Tier 1	lbe	Tier 2	os	Tier 2	stq	Tier 1	zh-yue	Tier 2
										zu	Tier 2

Table 15: Quality categorisations for all non-English Wikipedias.

E Heuristics per Tier

Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.	Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.	Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.
de	317.00	176.00	6.88	11.46	0.30	sah	83.00	61.00	5.62	7.77	0.33	ff	43.00	33.00	4.88	9.83	0.33
ru	251.00	152.00	6.71	10.38	0.31	gd	92.00	59.00	5.59	8.50	0.22	jam	55.50	40.00	5.04	17.38	0.33
es	259.00	138.00	6.44	11.12	0.25	sd	107.00	70.00	5.81	15.67	0.25	kbp	292.00	146.00	6.56	26.62	0.38
it	168.00	107.00	6.34	8.00	0.21	yi	119.00	75.00	5.89	9.56	0.31	wo	63.00	44.00	5.25	6.00	0.25
pl	135.00	90.00	6.11	8.04	0.23	li	184.00	107.00	6.26	14.29	0.33	kbd	81.00	59.00	5.50	7.83	0.33
ja	462.00	203.00	6.93	11.87	0.27	fo	86.00	52.00	5.41	7.14	0.22	nqo	200.50	110.50	6.23	12.42	0.29
pt	149.00	90.00	6.09	10.43	0.27	as	358.00	212.00	7.19	14.71	0.27	bi	24.00	18.00	4.07	5.67	0.33
ca	248.00	132.00	6.43	12.86	0.26	wa	112.00	71.00	5.80	7.92	0.33	tet	82.00	58.00	5.56	7.00	0.33
ko	114.00	76.00	5.86	5.75	0.20	hyw	237.00	155.00	6.78	11.65	0.33	roa-rup	46.00	35.00	4.93	8.98	0.33
fr	236.00	129.00	6.45	8.47	0.24	sn	90.00	62.00	5.67	7.33	0.42	jbo	66.00	39.00	4.99	11.40	0.00
no	137.00	85.00	6.02	7.57	0.20	co	75.00	51.00	5.39	6.67	0.25	fj	27.00	22.00	4.37	7.00	0.33
tr	96.00	68.00	5.79	6.14	0.18	so	90.00	64.00	5.75	9.39	0.25	guw	174.00	93.00	6.09	15.54	0.33
cs	242.00	154.00	6.78	9.29	0.20	vls	161.00	97.00	6.14	11.16	0.30	cu	38.00	30.00	4.76	6.20	0.00
eo	130.00	80.00	5.90	8.69	0.25	nds-nl	166.00	101.00	6.21	11.29	0.30	rmy	24.00	21.00	4.33	7.00	0.40
he	393.00	237.00	7.31	12.50	0.28	sc	117.00	74.00	5.79	14.80	0.33	trv	177.00	95.00	6.01	8.93	0.39
da	146.00	92.00	6.14	9.44	0.25	vep	238.00	153.00	6.65	9.20	0.33	mad	74.00	52.00	5.42	10.33	0.33
bg	187.00	112.00	6.31	10.66	0.27	kw	54.00	39.00	5.10	8.00	0.25	sm	50.00	33.00	4.81	11.00	0.33
et	102.00	70.00	5.77	6.73	0.28	kab	51.00	37.00	4.95	7.86	0.25	gcr	152.00	86.00	5.99	14.44	0.31
el	280.00	162.00	6.79	13.43	0.27	rw	125.00	88.00	6.17	10.00	0.25	pcm	218.00	120.00	6.43	18.27	0.41
hr	157.00	100.00	6.15	9.30	0.26	fiu-vro	37.00	28.00	4.57	4.11	0.25	gpe	221.00	116.00	6.37	12.82	0.23
lt	128.00	83.00	5.89	9.00	0.30	gv	99.00	64.00	5.68	8.67	0.20	pih	31.00	23.00	4.32	6.11	0.33
gl	180.00	108.00	6.29	9.20	0.21	mt	462.00	210.00	6.67	18.10	0.30	kcg	75.00	54.00	5.58	13.00	0.25
sl	139.00	90.00	6.03	7.38	0.22	frp	54.00	43.00	5.26	4.67	0.24	ss	82.00	58.00	5.55	8.11	0.35
nn	122.00	80.00	6.00	9.00	0.25	pcd	93.00	65.00	5.67	5.67	0.24	gur	164.00	84.00	5.88	12.33	0.26
ta	116.00	85.00	6.10	7.60	0.23	gn	52.00	39.00	5.11	7.10	0.27	ee	36.00	29.00	4.69	7.00	0.33
th	262.00	142.00	6.69	13.53	0.03	csb	46.00	37.00	5.05	7.07	0.26	chy	9.00	8.00	2.88	2.00	0.25
bn	227.00	138.00	6.67	9.00	0.20	smn	89.00	61.00	5.52	4.73	0.20	ik	10.00	10.00	3.22	3.00	0.33
lv	135.00	88.00	6.01	9.15	0.26	ay	108.00	72.00	5.86	2.96	0.27	fon	92.00	58.00	5.52	8.58	0.32
af	162.00	95.00	5.97	8.44	0.20	lez	92.00	67.00	5.79	5.56	0.25	ady	41.00	33.00	4.87	6.00	0.33
br	89.00	59.00	5.57	7.33	0.29	olo	60.00	42.00	5.14	4.90	0.27	guc	100.00	68.00	5.80	14.59	0.38
ml	147.00	105.00	6.35	9.16	0.24	mwl	169.50	100.00	6.08	14.70	0.33	fat	241.00	125.00	6.39	18.95	0.33
be-x-old	101.00	71.00	5.84	8.37	0.30	lfn	176.00	92.00	5.95	17.40	0.35	pwn	198.00	89.00	5.75	16.45	0.36
nds	166.00	101.00	6.21	11.29	0.30	kaa	64.00	49.00	5.35	9.74	0.32	din	206.00	115.00	6.24	18.54	0.33
lb	106.00	67.00	5.72	7.50	0.25	stq	80.00	55.00	5.50	10.33	0.33	ti	27.00	23.00	4.38	7.00	0.33
ga	56.00	42.00	5.20	6.86	0.20	ang	51.00	39.00	5.11	9.00	0.33	kl	61.00	45.00	5.26	8.79	0.33
is	96.00	68.00	5.80	10.07	0.27	fur	68.00	51.00	5.47	7.83	0.30	dz	1008.50	214.50	4.82	86.37	0.54
fy	231.00	127.00	6.39	12.07	0.27	ext	90.00	63.00	5.70	13.00	0.33	fi	141.00	97.00	6.22	8.82	0.23
pa	186.00	111.00	6.39	13.54	0.24	tw	218.00	116.00	6.32	16.24	0.31	sg	11.00	10.00	3.26	2.33	0.11
tl	149.00	84.00	5.91	12.50	0.25	lad	117.00	74.00	5.84	9.14	0.20	ts	76.00	52.00	5.40	11.52	0.33
an	116.00	70.00	5.74	8.60	0.30	gom	319.00	207.00	7.00	13.00	0.31	bm	16.00	13.00	3.58	2.44	0.14
wuu	65.00	48.00	5.41	42.00	1.00	pap	102.00	63.00	5.61	11.00	0.29	ny	68.00	49.00	5.42	12.38	0.33
sco	108.00	66.00	5.68	7.18	0.18	tyv	238.00	143.00	6.57	9.87	0.33	ltg	60.00	44.00	5.21	6.86	0.33
ha	182.50	108.00	6.34	13.42	0.27	ln	33.00	25.00	4.46	5.08	0.25	om	75.50	56.00	5.55	15.55	0.33
ne	62.00	48.00	5.44	5.40	0.17	ksh	117.00	81.00	6.03	10.77	0.33	inh	47.00	37.00	5.03	5.90	0.20
kn	263.00	177.00	6.95	12.67	0.30	bxr	89.00	63.00	5.69	8.97	0.25	dv	78.00	58.00	5.63	10.97	0.37
als	296.00	165.00	6.76	11.42	0.30	blk	141.00	100.00	6.21	38.05	0.41	dsb	69.00	50.00	5.38	5.83	0.21
bar	106.00	72.00	5.86	6.79	0.25	pdc	37.00	28.50	4.68	6.00	0.33	skr	259.00	145.00	6.65	15.85	0.33
ig	310.00	154.00	6.57	16.87	0.29	to	61.00	42.00	5.11	10.40	0.33	sat	333.00	154.00	6.62	13.27	0.24
si	131.00	91.00	6.20	11.76	0.29	tcy	204.00	136.00	6.63	9.42	0.25	km	146.50	88.00	5.81	15.41	0.25
frr	53.00	39.00	5.08	6.67	0.20	arc	22.00	20.00	4.25	7.67	0.33	mn	158.00	104.00	6.34	8.62	0.21
ps	165.00	95.00	6.07	17.28	0.33	mnw	49.00	36.00	5.05	18.57	0.29						

Table 16: Heuristic values for Tier 1.

Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.	Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.	Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.
nl	52.00	38.00	5.06	10.00	0.33	or	166.00	104.00	6.32	8.32	0.20	xh	75.00	57.00	5.56	14.73	0.40
uk	161.00	105.00	6.30	7.67	0.28	bcl	113.00	69.00	5.73	14.00	0.29	nia	24.00	21.00	4.30	4.40	0.20
fa	71.00	48.00	5.34	4.39	0.18	ilo	76.00	46.00	5.11	8.07	0.19	nov	56.00	40.00	5.09	6.00	0.33
id	98.00	66.00	5.64	8.75	0.24	am	27.00	21.00	4.33	4.50	0.25	ki	19.00	16.00	3.93	5.33	0.33
hu	187.00	115.00	6.30	7.54	0.21	sa	72.00	50.00	5.38	3.14	0.14	tn	119.00	72.00	5.84	17.67	0.33
ro	74.00	48.00	5.29	6.15	0.20	zu	16.00	15.00	3.81	2.67	0.20	kg	16.00	11.00	3.38	5.36	0.33
hy	139.00	92.00	6.17	7.64	0.23	hif	39.00	29.00	4.67	5.17	0.18	lbe	18.00	17.00	4.00	3.67	0.33
cy	146.00	91.00	6.12	6.79	0.21	mrj	34.00	29.00	4.69	5.36	0.33	ami	131.50	61.00	5.34	9.30	0.35
uz	69.00	48.00	5.38	6.56	0.25	ary	123.00	71.00	5.81	7.48	0.24	alt	305.00	187.00	6.95	8.18	0.33
sk	81.00	57.00	5.57	6.40	0.20	roa-tara	21.00	18.00	4.12	4.00	0.25	got	24.00	20.00	4.19	4.56	0.25
be	101.00	71.00	5.84	8.37	0.30	pam	61.00	46.00	5.33	5.00	0.15	rn	21.00	18.00	3.94	4.29	0.33
ur	76.00	51.00	5.44	5.08	0.14	bh	74.00	48.00	5.23	7.63	0.20	ch	24.00	19.00	4.17	5.50	0.25
az	112.00	79.00	5.99	7.50	0.25	myv	43.00	33.00	4.93	3.83	0.25	pnt	68.00	51.00	5.44	4.75	0.21
ka	92.00	66.00	5.74	7.33	0.20	se	18.00	16.00	3.88	3.00	0.20	iu	14.00	12.00	3.46	4.26	0.33
hi	75.00	46.00	5.31	7.36	0.18	bo	122.00	47.00	3.98	22.33	0.60	st	20.00	18.00	4.17	15.00	0.44
mk	198.00	118.00	6.37	9.93	0.25	tk	19.00	17.00	4.04	3.67	0.20	tpi	17.00	14.00	3.77	4.33	0.33
la	71.00	54.00	5.54	5.15	0.15	zea	111.00	69.00	5.70	7.64	0.29	anp	71.00	50.00	5.35	7.00	0.18
zh-yue	48.00	34.00	4.91	7.20	0.22	udm	38.00	29.00	4.66	5.29	0.33	krc	31.00	27.00	4.70	3.20	0.20
ast	176.00	109.00	6.23	8.13	0.22	kv	65.00	44.00	5.08	4.70	0.25	awa	21.00	17.00	3.95	6.00	0.33
my	60.00	42.00	5.18	11.20	0.20	nrm	84.00	54.00	5.32	4.67	0.23	av	41.00	34.00	4.93	3.80	0.22
sq	80.00	57.00	5.49	7.32	0.20	ks	24.00	20.00	4.25	4.20	0.17	nah	18.00	15.00	3.79	1.82	0.14
oc	70.00	50.00	5.43	7.00	0.20	lo	87.00	61.00	5.60	12.02	0.16	ug	82.00	61.00	5.67	13.60	0.33
sw	62.00	42.00	5.14	5.80	0.17	rm	56.00	41.00	5.15	7.00	0.25	rue	56.00	43.00	5.21	7.25	0.25
ky	115.00	82.00	5.93	10.57	0.22	dty	39.00	30.00	4.74	3.40	0.15	mhr	60.00	39.00	4.93	6.67	0.30
jv	52.00	38.00	5.06	5.90	0.19	lg	61.00	45.00	5.18	7.33	0.33	hsb	116.00	83.00	5.97	6.70	0.28
ba	174.00	107.00	6.31	7.89	0.25	za	22.00	17.00	3.97	3.67	0.33	os	19.00	16.00	3.93	3.60	0.20
io	63.00	43.00	5.18	10.33	0.33	gag	27.00	24.00	4.50	2.10	0.20	ckb	66.00	46.00	5.32	5.29	0.17
vo	51.00	36.00	4.98	5.18	0.30	szy	181.00	88.00	5.78	8.04	0.33	pnb	89.00	51.00	5.33	4.91	0.18
scn	29.00	23.00	4.45	6.33	0.33	tay	125.00	69.00	5.70	5.60	0.43	bs	182.00	109.00	6.27	6.07	0.22
qu	65.00	46.00	5.26	3.40	0.25	atj	33.00	26.00	4.61	5.00	0.20						
xmf	72.00	53.00	5.48	6.80	0.18	shi	87.00	51.00	5.22	9.00	0.22						

Table 17: Heuristic values for Tier 2.

Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.	Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.
sv	42.00	32.00	4.85	5.29	0.17	bjn	19.00	16.00	3.92	17.00	1.00
arz	51.00	36.00	4.98	2.46	0.18	hak	70.00	42.00	4.57	6.43	0.17
vi	40.00	32.00	4.93	4.33	0.14	nso	24.00	21.00	4.33	7.00	0.33
sr	70.00	51.00	5.48	3.76	0.15	gan	28.00	23.00	4.37	4.33	0.17
sh	70.00	51.00	5.48	3.76	0.15	tly	25.00	18.00	3.91	6.25	0.20
eu	51.00	41.00	5.19	4.50	0.20	mdf	53.00	39.00	5.03	2.09	0.15
kk	99.00	76.00	5.97	6.82	0.20	koi	52.00	40.00	4.92	4.21	0.25
tg	51.00	39.00	5.09	3.45	0.12	cbk-zam	64.00	46.00	5.27	2.80	0.12
lmo	55.00	38.00	5.01	6.36	0.20	pfl	90.00	57.00	5.53	3.92	0.30
vec	37.00	27.00	4.57	4.10	0.13	haw	24.00	18.00	4.02	7.00	0.33
ht	49.00	37.00	5.03	2.57	0.11	ty	17.00	12.00	3.50	5.00	0.33
pms	62.00	41.00	5.17	14.67	0.33	srn	56.00	38.00	5.02	9.05	0.20
su	31.00	23.00	4.47	11.73	0.29	chr	17.00	16.00	3.91	1.75	0.14
szl	36.00	29.00	4.71	8.50	0.25	ve	26.00	23.00	4.46	5.67	0.33
diq	47.00	36.00	4.95	2.71	0.14	cr	9.00	8.00	3.00	2.33	0.33
yo	10.00	9.00	3.12	1.60	0.20	xal	14.00	14.00	3.73	2.64	0.17
ia	30.00	22.00	4.32	9.00	0.33	ab	16.00	16.00	4.00	0.03	0.00
gu	93.00	63.00	5.69	19.00	0.33	glk	31.00	25.00	4.55	6.55	0.33
bpy	112.00	78.00	5.87	6.07	0.27	dag	75.00	52.00	5.27	2.25	0.07
mzn	53.00	40.00	5.19	5.89	0.17	cdo	21.00	16.00	3.73	6.00	0.33
bat-smg	21.00	19.00	4.20	5.67	0.33	tum	37.00	23.00	4.28	5.40	0.17
nap	39.00	18.00	3.76	5.14	0.14	ban	76.00	50.00	5.43	5.08	0.15
gor	29.00	18.00	4.03	3.71	0.14	nv	68.00	48.00	5.27	11.62	0.50
mai	49.00	36.00	5.11	3.30	0.11	crh	21.00	19.00	4.20	6.33	0.33
map-bms	52.00	38.00	5.06	5.90	0.19	ku	36.00	27.00	4.55	3.75	0.20
shn	103.00	53.00	5.48	8.40	0.20	te	341.00	211.00	7.02	10.42	0.25
eml	14.00	9.00	2.77	1.67	0.00	mr	89.00	58.00	5.45	5.50	0.25
ace	29.00	19.00	4.11	4.83	0.17	mg	58.00	35.00	4.97	3.53	0.20
ie	39.00	27.00	4.57	10.00	0.33	ms	47.00	32.00	4.79	4.70	0.18
lij	36.00	31.00	4.88	1.00	0.04						

Table 18: Heuristic values for Tier 3.

Wiki-Id	Len.	Uniq.	Ent.	W/L	Pct.
war	40.00	29.00	4.65	7.20	0.20
ce	88.00	64.00	5.65	4.76	0.18
tt	108.00	72.00	5.79	6.17	0.21
azb	49.00	38.00	5.10	4.67	0.20
min	57.00	47.00	5.46	5.64	0.25
lld	39.00	28.00	4.68	3.40	0.10
new	69.00	47.00	5.33	3.44	0.24
cv	105.00	72.00	5.82	5.32	0.18
avk	162.00	76.00	5.27	7.60	0.44
mni	27.00	23.00	4.46	4.00	0.14
mi	91.00	54.00	5.41	9.11	0.23
pag	58.00	42.00	5.17	3.38	0.06
pi	15.00	14.00	3.77	0.83	0.08
bug	19.00	12.00	3.38	3.33	0.17
ceb	88.00	46.00	5.11	10.08	0.27

Table 19: Heuristic values for Tier 4.