COMMON CORPUS: THE LARGEST COLLECTION OF ETHICAL DATA FOR LLM PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are pre-trained on large data from different sources and domains. These data most often contain trillions of tokens with large portions of copyrighted or proprietary content, which hinders the usage of such models under AI legislation. This raises the need for truly open pre-training data that is compliant with the data security regulations. In this paper, we introduce Common Corpus[1], the largest open dataset for LLM pre-training. The data assembled in Common Corpus are either uncopyrighted or under permissible licenses and amount to about two trillion tokens. The dataset contains a wide variety of languages, ranging from the high-resource European languages to some low-resource languages rarely represented in pre-training datasets. In addition, it includes a large portion of code data. The diversity of data sources in terms of covered domains and time periods opens up the paths for both research and entrepreneurial needs in diverse areas of knowledge. In this paper, we present the detailed provenance of data assembling and the details of dataset filtering and curation. We train two small language models on Common Corpus and find that the resulting model performs comparably to other models of their size, indicating that our dataset is suitable for multilingual pretraining. Common Corpus represents a key contribution to the ecosystem for open science research on large language models.

1 INTRODUCTION

Large Language Models have been defined by large amounts of training data. While there are several candidates for the first modern language model based on transformer architecture, including GPT-1 (Radford et al.), ULMFIT (Howard & Ruder, 2018), or Sentence Neuron (Radford et al., 2017), it is commonly acknowledged that "large" models start with GPT-3 (Brown et al., 2020). Requiring a corpus of 300 billion tokens, GPT-3 introduced a standard training data pipeline shared by nearly all language models to date: large-scale processing of web datasets (45 TB of compressed source data from Common Crawl) and additional digitized sources (Books3). Until 2025, LLM training data has grown on a logarithmic curve. The latest generation of publicly documented language models including DeepSeek v3 (Liu et al., 2025), Gemma 3 (Kamath et al., 2025), Llama 4 (Meta, 2025) or Qwen 3 (Yang et al., 2025) have been trained on 14-36 trillion tokens. Even the recently introduced sub-category of *Small language models* (Wang et al., 2025) relies on large amounts of training data to fit scaling laws: Qwen 3 0.6B was trained on 36 trillion tokens, which is a 3,000 times multiple of the original Chinchilla laws (Hoffmann et al., 2022).

As data curation became a major concern, the collection, maintenance, processing, and filtering of data became one of the main costs in language model training, not to mention even larger hidden costs: negative externalities affecting competing markets, the digital commons, and society at large.

While data scraped from the web is publicly available, it is not always in the public domain. Most web data does not have sufficient metadata to determine whether it is permissively licensed. NLP practitioners have relied on the protection of fair use, claiming that the transformative nature of the use of the data allows them to use this data to train language models. There are increasingly more legal challenges to the use of this data. The New York Times sued OpenAI for copyright

[1]The data will be made publicly available as a dataset on Hugging Face. We include a small sample from the dataset in the supplementary material.

infringement, alleging that OpenAI trained their models on NYT articles (Roth, 2023; Pope, 2024). Due to concerns about indirect commercial exploitation, many rightholders have implemented either hard technical measures or legal provisions against model training. In 2024, it was estimated that for Terms of Service crawling restrictions, a full 45% of C4 is now restricted (Longpre et al., 2024b) and 5% is fully blocked for scraping with a disproportionate impact over quality sources (Longpre et al., 2024b). Restrictions not only affect LLM pre-training but also the quality of search engine indexation and a variety of research projects analyzing and collecting content at scale. Even projects dedicated to knowledge access have faced significant pressure from AI crawlers and implemented protections that negatively impact access and user experience.

Legal uncertainties have significantly impeded the development of open science research on LLMs. Previously reproducible research artifacts have been removed or taken down, impacting pre-training data, continuous pre-trained models, and evaluation datasets. Books3, which has been used in datasets like the Pile (Gao et al., 2020), faced legal challenges (Brittain, 2023), and the original dataset was ultimately removed due to a DMCA takedown (Van der Sar, 2023). The LAION dataset was demonstrated to contain CSAM (Birhane et al., 2021; Thiel, 2023), and taken down (LAION, 2023), and then re-released once suspected CSAM was removed (LAION, 2024). The Dutch model GEITje was taken down (Rijgersberg, 2025), due to complaints about training on the Dutch Gigacorpus, in order to avoid legal disputes. Finally, the widely used benchmark, the Mathematics Aptitude Test of Heuristics (MATH) dataset (Hendrycks et al., 2021), was removed from Hugging Face via a DMCA takedown. All of these artifacts, which were released to further open development and evaluation of language models, were removed suddenly, making previous work unreplicable. These takedowns and legal challenges also represent a sizeable loss of investment for developers, who are often independent or small research organizations.

In part as a reaction to the use of publicly available but not permissively licensed data, web text is also becoming harder to acquire and use. In an analysis of popular datasets such as C4 (Raffel et al., 2020), RefinedWeb (Penedo et al., 2023), and Dolma (Soldaini et al., 2024), Longpre et al. (2024c) found that just in the last year, 5% of all tokens in C4 now have restricted use, with a disproportionate number of those tokens coming from the best-maintained, most critical sources. This is largely due to changes in content owners' and hosts' preferences, which are changing to no longer allow scraping, especially for the purposes of training AI models.

Since 2024, several initiatives have emerged to collect open data in English with clear licensing. This includes: C4C, Open License Corpus, a 228 billion token corpus from a mix of public domain texts and open source code under free licenses (Min et al., 2024), KL3M a 1.2 trillion tokens corpus of administrative texts and structured data mostly from the US federal public domain (Bommarito et al., 2025), Common Pile, a data collection of 1 trillion tokens from a variety of recent sources, including a filtered common crawl (Creative Commons Common Crawl) (Kandpal & Raffel, 2025). All these projects are monolingual, restricting in effect the reach of language models to the English-speaking audience. In contrast, the most ambitious multilingual collection of permissive content pre-dates Large Language Models: C4C (2016), containing 12 million web pages in more than 50 languages filtered by Creative Commons Licenses (Habernal et al., 2016).

Common Corpus has grown to become the largest fully open pre-training dataset at about **2 trillion tokens** and the only one in its size range having high multilingual diversity. Through this release, we show that open LLM research and development is possible while meeting legal and regulatory requirements — in compliance with even the strictest AI regulations, such as in the European Union. In this paper, we detail the composition of Common Corpus and the entire process of data collection and curation, and license clearing. Despite its size, Common Corpus is still far from covering the entire range of available resources: we attribute this discrepancy to an *open data paradox* as major sources of open content are paradoxically little visible online and even more so in the leading pre-training sources. By describing the unique challenges coming with the aggregation of large open source, we aim to inspire further initiatives. We also train two small language models on our dataset and find that it offers comparable performance to existing multilingual models.

2 About Common Corpus

When talking about Common Corpus data, we use the word **"open"** in the strongest possible sense. Not only is the data available, but we also provide essential details about the data provenance, data

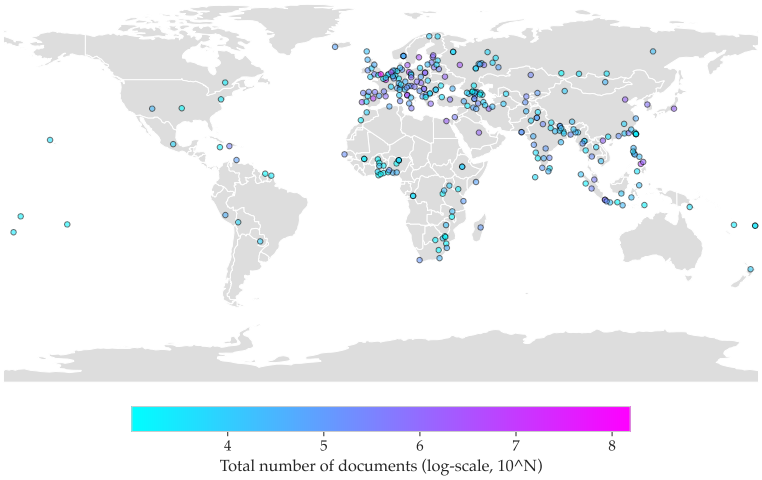


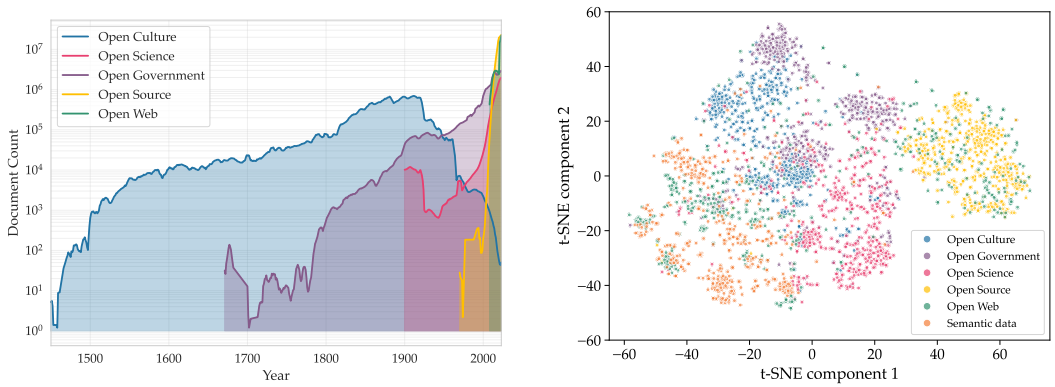
Figure 1: A schematic world map of languages in Common Corpus with a log-scaled distribution of document counts. For each language, we chose a city that is located in the region where this language is most specific to. To avoid outliers, we show only languages with 1000+ documents.

processing, and important information about the contents of each dataset. The Open Source Initiative has also defined open-source AI in terms of openness of use, where open means that use is permitted for "any purpose and without having to ask for permission" (Open Source Initiative, 2024). To achieve this, models must be trained on datasets that are free from copyright or other legal limitations. This is currently a limitation of existing open datasets for training LLMs.

Common Corpus, therefore, provides valuable training tokens that will not be subject to the same restrictions. Additionally, the data in Common Corpus are different from other corpora, primarily composed of web text. Common Corpus contains multilingual data in a variety of high- and low-resource languages (see Figure 1 for language distribution), covering diverse genres, time periods, and domains (in Section 3, we detail each part of the dataset). Therefore, Common Corpus contributes to data diversity in the open pre-training data ecosystem. This is important for developing powerful and generalizable model performance. Common Corpus can be used on its own or in conjunction with existing open datasets, according to one's needs and the desired use case of a language model.

Common Corpus was developed with consideration for ongoing conversations about best practices for open-source LLM development (The AI Alliance, 2024; Longpre et al., 2024a; Duprieu & Berkouk, 2024; Baack et al., 2025). We highlight our adherence to the best practices that were suggested by Baack et al. (2025):

- **Provide useful documentation.** We provide information about dataset provenance and processing (Sections 3 and 4) and share key statistics to help potential users understand the applications of the dataset. Dataset documentation improves reproducibility, helps prevent misuse, and aids downstream users to best utilize the dataset (Longpre et al., 2024a).
- **Follow and record preference signals.** In the metadata, we include the source URL and license information for the vast majority of the corpus.
- **Increase diversity and involve local communities to identify relevant data sources.** This dataset includes data from a variety of languages, coming from high-quality sources, and the multilingual part was never machine-translated.
- **Share advancements to foster reciprocity and give back.** In addition to the dataset, we release many of the tools we developed in order to create the final dataset (Section 4).
- **Do not use openly licensed data without regard for its quality or fitness for purpose.** In particular, for the dataset in the public domain, we engage in extensive OCR correction and toxicity filtering in order to bring datasets up to standard (Section 4).
- **Do not capture highly sensitive data.** We remove personally identifiable information from our datasets (Section 4).



(a) A timeline of the main collections with their numbers of documents in the Common Corpus.

(b) A two-component t-SNE visualization of a subset of the Common Corpus.

Figure 2: Temporal and semantic overview of the Common Corpus collections.

Table 1: Dataset composition of Common Corpus. For each collection, we report the total number of documents, words (whitespace-separated), and tokens.

Dataset	Documents	Words	Tokens
Open Government	74,727,536	257,233,670,261	406,581,454,455
Open Culture	93,156,602	549,608,763,966	885,982,490,090
Open Science	19,220,942	147,305,783,453	281,193,563,789
Open Code	202,765,051	77,669,169,092	283,227,402,898
Open Web	96,165,348	33,208,509,065	73,217,485,489
Semantic data	30,072,707	23,284,201,782	67,958,671,827
Other	925,462	328,160,421	486,099,734
Total	517,033,648	1,088,638,258,040	1,998,647,168,282

Common Corpus aims to support the pre-training of fully open and auditable LLMs by making it legal to release the source even without the provision of fair use. It has been used to create a wider range of language model artifacts, including multimodal datasets, classifiers, synthetic datasets, and benchmarks. Beyond the main dataset, Common Corpus works as an open science infrastructure dedicated to the entire lifecycle of language models. As defined by UNESCO, it is a shared research infrastructure that is needed to support open science and serve the needs of different communities (Unesco, 2021). We argue this is the first point in time where there has been sufficient knowledge and infrastructure to collect and clean a dataset on this scale, which meets the legal and ethical criteria we have outlined.

2.1 Composition

Common Corpus is available on HuggingFace as an aggregation of 10,000 parquet files and is composed of six collections: Open Government, Open Culture, Open Science, Open Web, Open Code, and Open Semantic. In total, the number of tokens in Common Corpus is **1,998,647,168,282**. The token counts[2] in each collection are listed in Table 1. We visualize the timeline of the collected documents and embeddings of a subsample in Figure 2. Each collection is composed of multiple datasets, for which we provide details about provenance and other key information in the corresponding subsections. Each data object contains a license, language(s), a collection/domain of specialization, and other metadata, allowing one to filter out a desired subset.

Common Corpus is multilingual (see Figure 1,Table 2, and Appendix C). Among many others, the top nine languages constitute at least 10B tokens each[3]. Some of the issues faced in making open datasets

[2]We report token counts in terms of the tokenizer trained on a representative subsample of Common Corpus.

[3]Language distribution was computed using the fastText language identification model.

Table 2: Token counts for the ten most represented languages in Common Corpus.

Language	Tokens
English	968,757,721,747
French	275,358,437,630
German	112,127,458,251
Spanish	46,514,142,421
Latin	36,031,591,540
Italian	24,681,637,575
Polish	12,146,688,669
Greek	11,376,498,056
Portuguese	10,262,747,943
Russian	9,439,453,633

Table 3: Token counts for the ten most common licenses in Common Corpus.

License type	Tokens
Public Domain	1,138,508,375,958
CC-By	287,749,264,457
MIT	142,694,227,607
CC-By-SA	74,768,060,836
Apache-2.0	68,750,977,037
BSD-3-Clause	18,483,944,333
Open license	10,432,513,767
BSD-2-Clause	5,497,145,480
CC-BY-4.0	2,110,966,243
CC0-1.0	1,877,206,195

for LLMs have been raised above, but all of these problems are much worse for languages other than English. Even in relatively high-resource languages like French, these problems are compounded by the fact that there is much less data available, and most tools generalize poorly to languages other than English. Additionally, Kreutzer et al. (2022) showed that many multilingual datasets contain a lot of low-quality or entirely unusable data. Many of the datasets they analyzed contained less than 50% of usable text, with 15 sources containing no usable data at all.

The majority of the data in Common Corpus is in the public domain (see Table 3). The license for each document is provided in the metadata, so the dataset can be easily filtered by license as desired.

3 Provenance

In this section, we present the details about collections that comprise the Common Corpus, accompanied by the information about the data sources and the main included languages in Appendix E.

3.1 Open Government

Open Government is a set of financial, legal, and administrative data in the public domain. In total, the dataset contains more than 406B tokens and comprises two main datasets: Finance Commons and Legal Commons. See Appendix E.1 for detailed data composition.

Finance Commons. This is the largest collection of financial documents in the public domain, comprising more than 14 billion words (more than 23 billion tokens). The documents come from a wide time range, all the way to 2024. Like many of our other datasets, Finance Commons is also multilingual. Most of the documents are in English, French, and German, but there are also texts in languages such as Romanian, Bulgarian, and Latvian. Additionally, this is a multimodal dataset. It includes more than 1.36 million original PDF documents from AMF and the WTO. The documents constitute a wide coverage of in-house layouts and formats produced by industrial and economic sectors. This makes this dataset ideal for developing the next generation of open-data multimodal models. One application for this dataset is to develop vision-language models (VLMs) for advanced document segmentation and processing. These documents also contain vast amounts of structured data, which is also a promising area of research that Finance Commons can help drive forward.

Legal Commons. This is a collection of legal and administrative datasets. The datasets come mostly from the EU and the US and cover a wide range of languages. These datasets are useful for developing language models with legal knowledge, as well as models that are ideal for document processing in official administrative applications.

3.2 Open Culture

Open Culture is an aggregation of vast cultural heritage datasets containing both monographs and periodicals for over 13 languages: French, English, German, Spanish, Portuguese, Italian, Dutch,

Luxembourgish, Danish, Swedish, Serbian, Czech, and Greek. There are also small portions of data in other languages, such as Arabic, Bengali, Latin, Persian, Russian, Sanskrit, and Urdu.

Composition. A large part of Open Culture is compiled from Collections As Data (CAD) — large dumps of texts, datasets, PDFs, and even raw XML output (METS/ALTO). CAD initiatives, thus, considerably simplify dataset aggregation and are a major contribution to the digital commons ecosystem. All other parts of Open Culture have been collected on a resource-by-resource basis using APIs and other standard retrieval methods whenever available. The largest extractions of this kind include Internet Archive (about 2 million monographs in multiple languages) and Delpher (50,000 Dutch monographs and periodicals filtered to match the Dutch copyright law for public domain). We managed to compile a large multilingual collection despite such challenges, as poor OCR quality, which we partly solved through the development of OCR correction tools (see Section 4), text segmentation issues, and sometimes irrecoverable deterioration of the original support. For the detailed dataset composition, refer to Appendix E.2.

Licenses. All Open Culture documents are in the public domain, which means their copyright has expired after a given term and there are no limitations on their reuse. For certain content, or in cases where we could not rely on the guarantee of established cultural heritage institutions, we implemented our own internal rights verification process. This process follows specific criteria, including author life and data object creation time, and takes into account that we only collected cultural heritage content from institutions based in the US or the EU (see the complete criteria list in Appendix F).

Value. Open Culture data is also rich from a cultural and stylistic standpoint and can be used to train multilingual language models with more diverse and creative writing styles. As LLMs are trained on extremely large corpora to maximize next-word prediction accuracy, LLM-generated text can often lack in personality and be boring or generic (Jones & Bergen, 2024). This feature of language models stands in contrast with one of their most common uses. In an analysis of WildChat (Zhao et al., 2024), a dataset of 1 million user interactions with ChatGPT, Longpre et al. (2024c) found that over 30% of user requests involved creative compositions such as fictional stories, role-play, or poetry generation. At the same time, creative writing is poorly represented among datasets used to train LLMs, which mainly comprise web text (Longpre et al., 2024c). Therefore, Open Culture contributes data that can be used to train models for creative writing without violating copyright law. In addition, as many of the Open Culture datasets are historical (coming from the 18th-19th centuries, or even earlier; see Figure 2a), this collection also enables the development of historical language models. The metadata includes document creation year, which enables researchers to develop language models with a cutoff of the training data creation date.

3.3 Open Science

The Open Science collection includes scientific papers and other documents (theses, book reviews, clinical trials, *etc*). Following the development of a global open access movement, these documents have been made increasingly available in open archives (preprints) or directly through open science publishers and infrastructure. Scientific content has become a primary focus of training data, due to its impact on reasoning capacities. Yet, the lack of licensing information has until now partly hindered reuse. The Semantic Scholar Open Research Corpus from Allen AI includes 81.1 million articles in English under an Open Data Commons Attribution License, allowing for the free reuse of the aggregated metadata while still acknowledging the remaining copyright of individual authors (Lo et al., 2020). The Pile incorporated data from arXiv and PubMed Central, also exclusively in English (Biderman et al., 2022). Finally, the BigScience project assembled several curated multilingual scientific datasets like the French HAL as part of the training data for Bloom (Scao et al., 2023).

The Open Science collection was made possible largely due to the recent development of OpenAlex[4], the largest open catalogue of scientific documents. OpenAlex maintains an expansive API search engine tracking detailed metadata for each indexed item, including the licensing, as well as a link to the original resource, which is generally in PDF format. We filtered OpenAlex on the three following licenses: CC-By, Public Domain/CC0, and CC-By-SA. The largest share of resources is available under CC-By, which is currently the recommended license by the Open Access definition. Open Science also includes smaller subsets, such as a direct extraction of arXiv articles available in CC-By

[4] https://openalex.org/

and some European-specific resources not currently well indexed on OpenAlex (the exact distribution of token counts can be found in Appendix E.3).

Due to the specificity of open scientific publishing, the Open Science collection has less linguistic diversity, with nearly 85% of documents currently available in English.

3.4 Open Code

The Open Code collection comprises code data under a vast variety of free licenses, which allows NLP practitioners to train models on public domain code for either coding applications or in order to improve certain model performance on natural language reasoning, world knowledge tasks, mathematics, and structured output tasks (Aryabumi et al., 2024; Petty et al., 2024; MA et al., 2024). The code data we use comes from the Stack v1 and v2 (Kocetkov et al., 2023; Lozhkov et al., 2024). The Stack v1 contains 6.4TB of data and covers 30 programming languages, while the Stack v2 is approximately ten times bigger at 67.5TB and covers over 600 programming languages. All the code data is made available with a direct link to the original resource on GitHub. In total, Open Code contains 283,227,402,898 tokens (see most common languages in Appendix E.4).

To prepare the collection, we ran a pipeline of varied filters. We first removed files that were not in our desired set of languages and formats according to their file extensions, including SVG files containing mostly encoded shapes, data storage formats: `csv`, `json`, `json5`, `jsonld`, and other file types with non-informative content, typically in small amounts: `python-traceback`, `unity3d-asset`, `numpy`, and `http`. We then filtered out the licenses to keep only permissible ones. To discard the low-quality data, we ran a series of manual filters described by Lozhkov et al. (2024). In addition to those, we removed files consisting of 75% or more of digits, which are mostly files containing raw numeric data. Before the filters, we also replaced sequences of `[\r]+\n` with `\n` and recalculated line lengths to avoid false positives by maximum line length.

3.5 Open Web

In accordance with the general focus of Common Corpus on curated content, the Open Web collection currently includes four major web sources:

Wikipedia and **Wikisource.** Wikimedia projects have always been major sources for language model training due to their reliability, extensive coverage, and textbook-like style. Despite this centrality, there is still a range of unresolved challenges with the most common versions available for training. The raw source of Wikimedia projects is made available in a specific *mediawiki* syntax, including a lot of project-specific models, tags, and conventions. The parsing of models is especially not straightforward, as they can either format existing text or remove or include external content (transclusion). As part of Wikimedia Enterprise, the Wikimedia Foundation created entirely new dumps from the rendered HTML sources, which in effect ensure that they include all the text made available to readers.

Youtube Commons. For YouTube Commons, we collected audio transcripts of 2,063,066 videos uploaded on YouTube under a standardized CC-By license.

StackExchange. This is a collection of user-generated forums and Q&A made available under the CC-By-SA license. We reused the version from The Pile (Biderman et al., 2022).

A major objective for the future work will be the integration of web archives filtered by permissive licenses. Since 2016, several projects have attempted to reidentify Creative Commons licenses from web archives at scale including C4C (multilingual) (Habernal et al., 2016) and more recently CCCC (from Allen AI, in English) and most recently Common Crawl Creative Commons Corpus (C5, for the first time multilingual)[5]. All these projects struggled with license identification. While license mentions are frequently normalized with a direct link or logo to Creative Commons, there is no guarantee they really concern the entire content: "a blog page contains many photos, and each photo is licensed under a different CC-license type, or a blog home page with many articles, and each article is licensed under a different CC-license type." (Habernal et al., 2016). We hope this limitation could be overcome by a combination of web domain curation and fine-grained curation and annotation by a language model.

[5] `https://huggingface.co/datasets/BramVanroy/CommonCrawl-CreativeCommons`

3.6 Open Semantic

Semantic data is the latest set added to Common Corpus and currently includes only one collection: Wikidata. First created in 2011, Wikidata hosts 100 million documented items and several billion factual statements encoded as RDF triples. It has grown to become a critical web infrastructure, used by Google for search disambiguation and currently embodying Tim Berners-Lee's ambitious vision for "a web of data". Despite the rising interest in mixed LLM/knowledge graph methods, Wikidata has hardly been used in language models. The largest initiative to date is Kelm, a collection of 15 million synthetic sentences generated by Google from English-speaking statements (Agarwal et al., 2021). A persistent challenge has been the exclusive availability of Wikidata dumps under formats optimized for data exchange rather than language model training.

Thanks to a collaboration with Wikimedia Deutschland, the entire set of Wikidata has been adapted in natural language and added to Common Corpus. This is to date the only available textual collection of Wikidata covering the entire range of 300 languages. Data processing involved the translation of items and properties into formal language sequences as simple natural language sequences, without textual synthesis: "Q41309 — P:27 — Q171150" becoming "Franz Liszt country of citizenship Kingdom of Hungary". Within each entry, we provide all the available translations as consecutive blocks separated by a newline, anticipating that this may contribute to language alignment.

4 Cleaning and Curation

In order to curate our dataset, we developed a number of custom tools to handle the issues unique to multilingual, historical, and OCRed data. We will release all of them under permissive licenses.

Text Segmentation. We developed **Segmentext**, a specialized language model for text segmentation (see example in Appendix G.1). Segmentext has been trained to be resilient to broken and unstructured texts with digitization artifacts and ill-recognized layout formats. Given the diversity of the training data, Segmentext should work correctly on diverse document formats in the main European languages.

OCR Correction. We developed **OCRonos** model based on Llama 3 8B (Grattafiori et al., 2024). OCRonos is versatile and supports the correction of OCR errors, cutting or merging of the wrong word, and overall broken text structures. The training data includes a highly diverse set of OCR-ed texts in multiple languages, mostly coming from uncorrected versions of Open Culture and Open Government. On highly deteriorated content, OCRonos can act as a synthetic rewriting tool rather than a strict correction tool. An example of OCRonos work is presented in Appendix G.3. OCRonos contributes to make challenging resources usable for LLM applications and, more broadly, search retrieval. It is especially fitting in situation where the original PDF sources is too damaged for correct OCR or even non-existent/complex to retrieve.

OCRonos is generally faithful to what the original material, provides sensible restitution of deteriorated text and will rarely rewrite correct words. On past experiments, a common issue with OCR correction has been language switching: due to the inherent noise in the input text, an LLM will transcribe in a different language or script. The issue has been especially observed in smaller generalist models like GPT-3.5 or Claude-Haiku. OCRonos largely mitigates this issue.

PII Removal. Personally Identifiable Information (PII), i.e., any information that can be used to distinguish or trace an individual's identity, is protected under legislation such as GDPR. Consequently, the new regulations put restrictions on LLM training data. In large open datasets, there is a staggering amount of personal data in widely used datasets, *e.g.*, large quantities of phone numbers in RedPajama, email addresses in S2ORC and peS2o, and IP addresses in the Stack (Elazar et al., 2024). To identify and replace PII, we use Microsoft's Presidio[6], an open-source state-of-the-art tool. With Presidio, we filtered out phone numbers, email addresses, IBANs, IP addresses, and URLs. With the base settings, Presidio identified on average 55-60% of texts that included phone numbers due to different possible number formats. By applying custom regular expression patterns that include most phone numbers, we increased this accuracy to 85%. Typical methods of handling PII include removing it, replacing it with tags, and partial anonymization. These transformations substantially alter the format of PII, which could undermine the model's understanding of the text or interfere with its ability to process text with real PII. Instead, we replace PII with fictitious but realistic values.

[6] https://microsoft.github.io/presidio/

Table 4: Multilingual benchmarking results. "Ours" refers to models pre-trained on Common Corpus.

Model	**Ours**	**Gemma 3**	**XGLM**	**BLOOM**	**Ours**	**Gemma 3**	**XGLM**	**OLMo**
Parameters	350M	270M	564M	560M	1.2B	1B	1.7B	1B
MultiBLiMP	0.774	0.762	0.711	0.683	0.797	0.799	0.710	0.699
XStoryCloze	0.509	0.533	0.537	0.532	0.526	0.594	0.569	0.517
XCOPA	0.533	0.544	0.550	0.541	0.541	0.593	0.574	0.518

Deduplication. Our early experiments showed a negligible rate of duplication, which we attribute to the initial data curation: large institutions are incentivized to avoid re-digitizing the same texts. We also filtered out duplicates based on PDF metadata and used deduplicated sources wherever possible.

Toxicity Detection. In addition to posing legal and regulatory issues, web data is a major source of harmful and biased content (Common Crawl was shown to contain sexual content, hate speech, and racial and gender biases (Luccioni & Viviano, 2021)) and often suffers from low-quality and machine-generated text (Dodge et al., 2021). Public Domain data, such as that in Open Culture, do not pose the same legal challenges but introduce new ones. Many texts there are historical periodicals and monographs from at least 80 years ago, while cultural norms have changed dramatically. Many of these texts, therefore, do not meet modern ethical standards. Training language models on these texts would lead to the reproduction and circulation of harmful language.

To address this, we developed a pipeline to filter the public domain training data. We identify documents containing harmful language and either remove it or synthetically rewrite the document without the harmful language. With this approach, we aim to mitigate some of the potential biases and harms in the dataset, while still leveraging the high-quality, diverse data for high model performance. We created a multilingual toxicity classifier, **Celadon**, a DeBERTa-v3-small model ($\sim$140M parameters), which we trained from scratch on 2M annotated samples. Celadon identifies toxic and harmful content along five dimensions: race and origin-based bias, gender and sexuality-based bias, religious bias, ability bias, and violence and abuse. We will release the model along with the training data.

5 Model Training

We train two models on Common Corpus: a 350M and a 1.2B model. The architecture is based on Llama. We train a custom Llama-style tokenizer with a vocabulary size of 65536 on a representative subset of Common Corpus. The 350M model is trained on a filtered subset of Common Corpus, comprised of approximately 1T tokens. The 1.2B model is trained on two epochs of Common Corpus. The models were trained for 2944 and 23040 H100 hours, respectively. We will release our models on Hugging Face. We will also release our full training pipeline under an Apache 2.0 license.

We evaluate our models on MultiBLiMP (Jumelet et al., 2025), XStoryCloze (Lin et al., 2022), and XCOPA (Ponti et al., 2020) (see Table 4 for aggregated scores and Appendix H for per-language details). All evaluations were run using the LM Evaluation Harness (Biderman et al., 2024). Our models perform comparably to models trained on closed or non-permissively licensed data, and show outstanding performance on MultiBLIMP, which has more languages compared to other benchmarks. This is especially notable for our 350M model, which we compare to bigger models; it also outperforms models from the 1B range, except for Gemma 3 1B. Our models stably outperform OLMo 1B, which was also pre-trained on a publicly released dataset.

6 Conclusion

Through the release of Common Corpus and this paper with thorough documentation of data collection and curation, we show that LLM development is possible while strictly adhering to the regulatory norms. While Common Corpus is only large enough to train small models currently, the tools and methods we used to identify and curate the data may be used to expand the amount of permissively licensed open data. We hope that Common Corpus will grow as a critical infrastructure for open science LLM research and development and inspire future initiatives in the open.

REFERENCES

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3554–3565, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.278. URL https://aclanthology.org/2021.naacl-main.278/.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training. *CoRR*, abs/2408.10914, 2024. URL https://doi.org/10.48550/arXiv.2408.10914.

Stefan Baack, Stella Biderman, Kasia Odrozek, Aviya Skowron, Ayah Bdeir, Jillian Bommarito, Jennifer Ding, Maximilian Gahntz, Paul Keller, Pierre-Carl Langlais, et al. Towards best practices for open datasets for llm training. *arXiv preprint arXiv:2501.08365*, 2025.

Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the Pile, January 2022. URL http://arxiv.org/abs/2201.07311. arXiv:2201.07311 [cs].

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Michael J. Bommarito, Jillian Bommarito, and Daniel Martin Katz. The KL3M Data Project: Copyright-Clean Training Resources for Large Language Models, April 2025. URL http://arxiv.org/abs/2504.07854. arXiv:2504.07854 [cs].

Blake Brittain. Authors sue meta, microsoft, bloomberg in latest ai copyright clash. *Reuters*, October 18 2023. URL https://www.reuters.com/legal/litigation/authors-sue-meta-microsoft-bloomberg-latest-ai-copyright-clash-2023-10-18/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on EU legislation. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6314–6322, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1636. URL https://aclanthology.org/P19-1636.

Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672, 2022. URL https://api.semanticscholar.org/CorpusID:250425961.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98.

Henri Duprieu and Nicolas Berkouk. Techniques d'audit des grands modèles de langage. Technical report, Commission Nationale Informatique et Libertés (CNIL), November 2024. URL https://hal.science/hal-04782667.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RvfPnOkPV4.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ivan Habernal, Omnia Zayed, and Iryna Gurevych. C4Corpus: Multilingual Web-size Corpus with Free License. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 914–922, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1146/.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset, 2022. URL https://arxiv.org/abs/2207.00220.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022. URL http://arxiv.org/abs/2203.15556. arXiv:2203.15556 [cs].

Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification, May 2018. URL http://arxiv.org/abs/1801.06146. arXiv:1801.06146 [cs].

Cameron Jones and Ben Bergen. Does GPT-4 pass the Turing test? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5183–5210, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.290. URL https://aclanthology.org/2024.naacl-long.290.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*, 2025.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Nikhil Kandpal and Colin Raffel. Position: The Most Expensive Part of an LLM should be its Training Data, April 2025. URL http://arxiv.org/abs/2504.12427. arXiv:2504.12427 [cs].

Denis Kocetkov, Raymond Li, Loubna Ben allal, Jia LI, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro Von Werra, and Harm de Vries. The stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=pxpbTdUEpD.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pp. 79–86, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

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, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. doi: 10.1162/tacl_a_00447. URL https://aclanthology.org/2022.tacl-1.4.

LAION. Safety review for laion 5b, December 19 2023. URL https://laion.ai/notes/laion-maintenance/.

LAION. Releasing re-laion 5b: Transparent iteration on laion-5b with additional safety fixes, August 30 2024. URL https://laion.ai/blog/relaion-5b/.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616/.

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, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The Semantic Scholar Open Research Corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault

(eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447/.

Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, Maribeth Rauh, Aviya Skowron, Bertie Vidgen, Laura Weidinger, Arvind Narayanan, Victor Sanh, David Ifeoluwa Adelani, Percy Liang, Rishi Bommasani, Peter Henderson, Sasha Luccioni, Yacine Jernite, and Luca Soldaini. The responsible foundation model development cheatsheet: A review of tools & resources. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=tH1dQH20eZ. Survey Certification.

Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Sandy Pentland. Consent in Crisis: The Rapid Decline of the AI Data Commons, July 2024b. URL http://arxiv.org/abs/2407.14933. arXiv:2407.14933 [cs].

Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad A. Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester James V. Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Sandy Pentland. Consent in Crisis: The Rapid Decline of the AI Data Commons. *CoRR*, abs/2407.14933, 2024c. URL https://doi.org/10.48550/arXiv.2407.14933.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. EDGAR-CORPUS: Billions of tokens make the world go round. In Udo Hahn, Veronique Hoste, and Amanda Stent (eds.), *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pp. 13–18, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.econlp-1.2. URL https://aclanthology.org/2021.econlp-1.2.

Eneldo Loza Mencía and Johannes Fürnkranz. Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain. In *Semantic Processing of Legal Texts*. Springer, 2010.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder 2 and the stack v2: The next generation, 2024. URL https://arxiv.org/abs/2402.19173.

Alexandra Luccioni and Joseph Viviano. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*,

pp. 182–189, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.24. URL https://aclanthology.org/2021.acl-short.24.

YINGWEI MA, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help LLMs reasoning? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KIPJKST4gw.

Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore, July 2024. URL http://arxiv.org/abs/2308.04430. arXiv:2308.04430 [cs].

Clemens Neudecker. An open corpus for named entity recognition in historic newspapers. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4348–4352, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1689.

Open Source Initiative. The open source ai definition – 1.0, 2024. URL https://opensource.org/ai/open-source-ai-definition. Accessed: 2024-11-20.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172, 2023.

Jackson Petty, Sjoerd van Steenkiste, and Tal Linzen. How does code pretraining affect language model task performance? In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=2sghJ1yYOr.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185/.

Audrey Pope. NYT v. OpenAI: The Times's About-Face. *Harvard Law Review Blog*, April 2024. URL https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to Generate Reviews and Discovering Sentiment, April 2017. URL http://arxiv.org/abs/1704.01444. arXiv:1704.01444 [cs].

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer . *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Edwin Rijgersberg. The end of geitje. GoingDutch.ai, 2025. URL https://goingdutch.ai/en/posts/geitje-takedown/. Accessed: 2025-02-20.

Emma Roth. New York Times sues OpenAI and Microsoft over copyright infringement. *The Verge*, December 2023. URL https://www.theverge.com/2023/12/27/24016212/new-york-times-openai-microsoft-lawsuit-copyright-infringement.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, et al. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL https://arxiv.org/abs/2211.05100.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint arXiv:2402.00159*, 2024. URL https://arxiv.org/pdf/2402.00159.

The AI Alliance. Dataset specification, 2024. URL https://the-ai-alliance.github.io/open-trusted-data-initiative/dataset-requirements/. Accessed: 2025-02-20.

David Thiel. Identifying and eliminating csam in generative ml training data and models. Technical report, Stanford Digital Repository, December 20 2023. URL https://purl.stanford.edu/kh752sm9123.

Unesco. Recommendation on Open Science, 2021. URL https://www.unesco.org/en/legal-affairs/recommendation-open-science.

Ernesto Van der Sar. Anti-piracy group takes prominent ai training dataset "books3" offline. *TorrentFreak*, August 16 2023. URL https://torrentfreak.com/anti-piracy-group-takes-prominent-ai-training-dataset-books3-offline-230816/.

Chengyu Wang, Taolin Zhang, Richang Hong, and Jun Huang. A Short Survey on Small Reasoning Models: Training, Inference, Applications and Research Directions, April 2025. URL https://arxiv.org/abs/2504.09100v1.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

A LLM USAGE STATEMENT

In the process of developing this work, we utilized LLMs for grammar correction and occasionally as a rewriting tool. In addition, we involved LLMs in the process of data visualization.

B LIMITATIONS

Common Corpus is far from collecting the whole range of available open data, which we described as the open data paradox. Therefore, the future collection of permissible data is highly encouraged by this work. Furthermore, the collected amount of data (2 trillion tokens), when used alone, as our own small language model family (see Section 5), is suitable for pre-training of models of limited size, while larger ones require significantly larger amounts of data. In addition, Common Corpus naturally does not contain data for instruction-tuning and any forms of specialized tasks. Therefore, it is not directly suitable for task-specific fine-tuning. However, due to the multilingual, temporal, and semantic diversity of data, Common Corpus opens the opportunities for the creation of ethical fine-tuning datasets.

In Section 4, we described the tools we used for the data curation, filtering, and editing. Even though we used these methods responsibly and mitigated many issues overlooked by the counterparts (*e.g.*, with toxicity detection), none of the curation methods could naturally facilitate a hundred-percent accuracy. However, some issues, like OCR errors, present considerable challenges to the models and might even account for better handling of typos in the future. We would also like to mention that each data object is accompanied by sufficient metadata, and, if desired, LLM practitioners are free to filter out collections that might contain potential issues (as described in Section 4).

C LANGUAGE DISTRIBUTION

In Table 5, we present the top-50 languages in Common Corpus by token count. The token counts are presented in terms of our BPE tokenizer used to train the models described in Section 5, which was trained on a representative subsample of Common Corpus. To verify that our tokenizer serves as a strong baseline for token counts, we show its fertility in Appendix D.

D TOKENIZER DETAILS

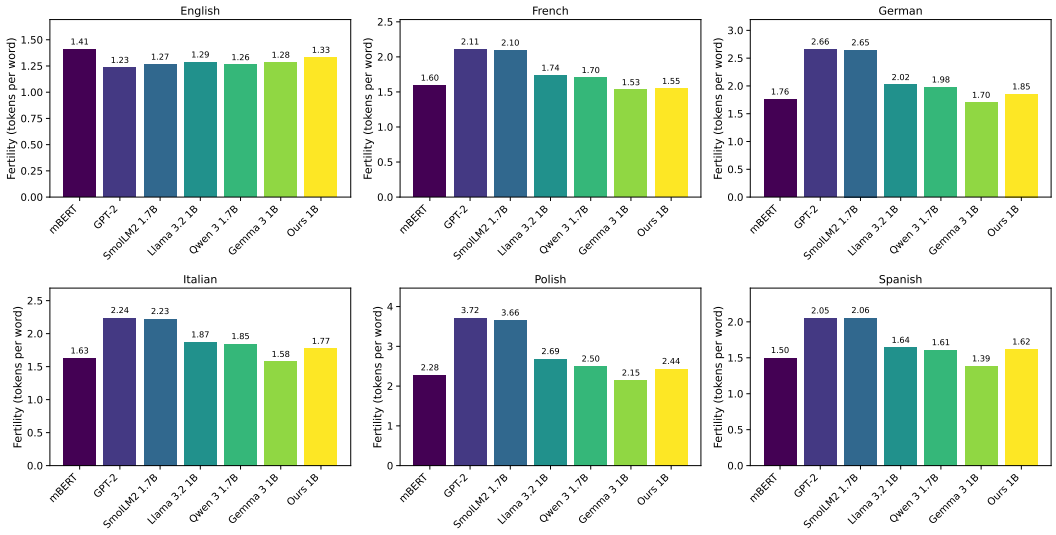


Figure 3: Comparing the fertility of our tokenizer (marked as "Ours 1B") and other language models for six languages. The data source for all languages is the devtest set of FLORES200 (Costa-jussà et al., 2022).

In Figure 3, we show how our tokenizer with a vocabulary of size 65,536, trained on a representative subsample of Common Corpus, compares to other language model tokenizers. Our tokenizer is outperformed only by Gemma 3, which has a tokenizer **four times larger**.

Table 5: Top-50 languages in Common Corpus by token count. Each language is presented with its number of documents, words, and tokens in the corpus. The rows are ordered by the token count.

Dataset	Documents	Words	Tokens
English	154,175,907	634,794,970,595	968,757,721,747
French	35,245,624	162,061,620,874	275,358,437,630
German	11,385,377	56,674,819,173	112,127,458,251
Spanish	6,530,094	26,215,767,271	46,514,142,421
Latin	2,367,110	16,189,444,325	36,031,591,540
Italian	3,804,052	13,207,129,356	24,681,637,575
Polish	2,640,613	5,086,555,167	12,146,688,669
Greek	844,122	3,796,018,483	11,376,498,056
Portuguese	1,756,922	5,234,373,473	10,262,747,943
Russian	2,762,818	3,222,919,854	9,439,453,633
Dutch	3,206,382	3,791,928,728	8,058,934,080
Danish	2,270,459	2,840,121,206	6,941,827,931
Slovak	683,174	2,320,831,403	5,148,967,838
Czech	946,534	1,966,829,784	4,798,558,092
Indonesian	1,023,361	1,660,129,567	4,381,878,823
Estonian	685,180	1,613,093,565	4,379,534,617
Hungarian	888,780	1,527,981,866	4,110,878,972
Swedish	3,250,289	1,782,642,556	4,014,927,806
Finnish	931,201	1,356,653,003	3,943,036,413
Maltese	480,491	1,421,372,608	3,646,102,921
Bulgarian	576,997	1,444,748,621	3,422,182,324
Lithuanian	539,906	1,192,586,834	3,097,400,907
Romanian	725,766	1,398,178,029	2,909,452,579
Japanese	1,409,956	204,698,439	2,738,872,745
Arabic	1,291,439	827,392,227	2,682,255,180
Slovenian	504,492	1,107,467,643	2,602,943,380
Latvian	364,503	966,005,785	2,579,350,119
Ukrainian	1,378,390	786,829,868	2,561,253,212
Croatian	443,669	984,577,685	2,400,762,140
Chinese	1,426,017	329,649,628	2,238,230,225
Haitian Creole	292,718	762,225,686	1,420,886,048
Turkish	572,633	380,551,359	1,206,732,266
Cebuano	6,123,694	598,016,557	1,080,404,897
Norwegian Nynorsk	1,026,745	405,911,677	1,036,914,970
Irish	139,815	371,175,864	958,688,647
Castilian	50,785	457,490,798	893,080,621
Serbian	698,037	292,399,267	822,519,328
Hebrew	343,406	227,869,199	763,029,488
Catalan	803,532	368,144,006	736,220,789
Korean	653,460	143,729,419	677,716,397
Persian	983,632	209,078,446	668,287,975
Vietnamese	1,296,789	267,190,147	593,203,801
Norwegian	628,798	252,133,513	561,694,623
Armenian	306,959	109,852,686	539,900,475
Hindi	210,984	127,830,222	463,832,707
Yiddish	59,158	122,100,684	463,526,449
Welsh	309,573	191,827,552	438,133,529
Occitan	288,195	195,380,666	357,267,635
Georgian	172,532	71,678,084	349,884,144
Basque	439,582	139,077,600	348,891,265

E PROVENANCE

E.1 OPEN GOVENMENT

In this section, we describe the provenance and present token counts and main languages for the two sub-collections of Open Government: Finance Commons and Legal Commons.

E.1.1 FINANCE COMMONS

Table 6: Finance Commons sources distribution with languages.

Dataset	Main Languages	Documents	Tokens
SEC	English	1,085,113	9,653,919,837
WTO	English, Spanish, French, and small partitions of others	772,508	2,835,007,015
AMF	French, English	595,397	9,823,755,281
TED EU Tenders	German, French, Polish, Spanish, Dutch, Czech, Romanian, English, Swedish, Italian, Bulgarian, Finnish, Latvian, Danish, Lithuanian, Croatian, Estonian, Hungarian, Portuguese, Slovenian, Slovak, Greek, Irish	137,837	650,396,761
GATT Library	English, French, Spanish, Catalan, Portuguese, German	67,596	224,526,628

The datasets that make up Finance Commons are presented in Table 6. Here, we also present the provenance details for each of the parts of Finance Commons:

- **Securities and Exchange Commission (SEC).** This dataset comprises the SEC annual reports (Form 10-K) for the years 1993 to 2024. Entries up to 2020 were compiled by Loukas et al. (2021). We added the reports from 2021-2024, which come from the EDGAR database[7], compiled using the EDGAR-Crawler toolkit[8].
- **World Trade Organization (WTO).** This dataset comprises documents from WTO's official Documents Online platform. The documents cover the years 1995 to 2024. Documents are available in three official languages: English, French, and Spanish. Some documents are available in other languages, *e.g.*, Chinese, Korean, Arabic, German, and Portuguese. Also released separately as WTO-PDF.
- **French Authority for Financial Market (AMF).** This is a dataset of documents from the French Authority for Financial Market, or the Autorité des marchés financiers[9] (AMF), which is an independent public authority that regulates the French market. The documents are primarily in French. Also released separately as AMF-PDF.
- **Tenders Electronic Daily (TED) EU Tenders.** This dataset is a collection of procurement notices published by the EU. The documents are published in the online version of the "Supplement to the Official Journal" of the EU[10], dedicated to European public procurement. The documents are mostly in German, with French, Polish, and Spanish making up relatively large portions of the remaining documents. There are also small portions of other languages (see details in Table 6).
- **General Agreement on Tariffs and Trade (GATT) Library.** This dataset comprises documents from GATT, which was an organization that promoted international commerce and the reduction of trade barriers among member states. Public documents were made

[7] https://www.sec.gov/search-filings/edgar-search-assistance/accessing-edgar-data

[8] https://github.com/nlpaueb/edgar-crawler

[9] https://www.amf-france.org/en/news-publications/publications/open-data

[10] https://ted.europa.eu/en/

Table 7: Legal Commons sources distribution with languages.

Dataset	Languages	Tokens
Caselaw Access Project	English	13,821,842,995
Court Listener	English	22,625,121,735
EUR-lex	Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish	65,044,763,781
Eurovoc	English, German, French, Croatian, Italian, Lithuanian, Portuguese, Finnish, Danish, Bulgarian, Dutch, Polish, Greek, Swedish, Hungarian, Czech, Spanish, Maltese, Latvian, Slovak, Slovenian, Romanian, Estonian, Arabic, Tigrinya, Farsi, Russian, Urdu, Serbian, Albanian, Kurdish, Pushto, Irish, Norwegian, Icelandic, Dari, Armenian, Japanese.	31,648,136,898
French open data	French	24,597,392,089
USPTO	English	200,509,900,178
UN Digital Library	Arabic, Chinese, English, French, Russian, Spanish	1,781,037,875
European Open Data	EU languages	7,098,502,579
OECD	English, French	584,969,458

available by the General Council of the WTO in 2006[11]. The documents span from January 1, 1946, to September 6, 1996. Most of the documents are in English, but there are also documents in French, Spanish, and other languages.

E.1.2 LEGAL COMMONS

Here, we present the provenance details for each of the parts of Legal Commons:

- **Europarl.** This dataset is a multilingual parallel corpus, drawn from the proceedings of the European Parliament[12]. It includes texts from 21 EU languages. It was originally compiled by Koehn (2005).
- **Caselaw Access Project.** This dataset consists of 6,773,632 legal cases, digitized from Harvard Law School Library's physical collection of American case law[13]. The dataset spans the years 1658 to 2020.
- **CourtListener.** This is a dataset[14] of opinions, oral arguments, judges, judicial financial records, and federal filings put together by the Free Law Project[15].
- **EUR-lex.** This is a dataset of 57,000 legislative documents from the EU[16]. It is based on the dataset by Loza Mencía & Fürnkranz (2010) and developed by Chalkidis et al. (2019). The

[11] https://www.wto.org/english/docs_e/gattdocs_e.htm
[12] https://www.statmt.org/europarl/
[13] https://case.law/
[14] https://www.courtlistener.com/help/api/bulk-data/
[15] https://free.law/contact
[16] https://eur-lex.europa.eu/

documents have also been annotated by the Publications Office of EU[17] with concepts from EuroVoc[18]. The dataset covers all 24 EU languages.

- **Eurovoc.** Eurovoc is a dataset containing 1,528,402 documents in 39 languages with associated EuroVoc labels. The documents come from Cellar[19], which is a data repository for the Publications Office of the European Union. This dataset was originally compiled by Sébastien Campion[20].
- **French Open Data.** This dataset comes from French administrative bodies' websites, for example, the French Directorate of Legal and Administrative Information (Direction de l'information légale et administrative[21]; DILA), which is a French public administrative entity that disseminates information about laws and their applications to the public.
- **USPTO.** This dataset comprises documents from the United States Patent and Trademark Office (USPTO), the federal agency that grants patents and registers trademarks. This dataset consists of actions from this agency from 2019 to 2022. It was originally published as part of the Pile of Law (Henderson et al., 2022)[22].
- **UN Digital Library.** This dataset comes from the UN Digital Library[23].
- **European Legal Dataset.** We also collect datasets from various EU websites, *e.g.*, Archives of the EU Institute[24] and the Council of the EU[25].
- **OECD**. These data come from the Organisation for Economic Co-operation and Development (OECD)[26].

E.2 Open Culture

Large portion of data in Open Culture part of the Common Corpus was built on top of the following collection-as-data initiatives:

- **Chronicle America**: about 100B words (150B tokens) of digitized US newspapers by the Library of Congress, made available as a raw text file.
- **Europeana**: about 21B tokens of digitized European newspapers through large-scale cross-national contributions and new digitizations.
- **Gallica**: about 85B words of digitized French newspapers and monographs made available on the open data portal of the French digitized library through entire dumps or API access[27].
- **Biblioteca**: about 15B words of digitized Spanish newspapers and monographs.

Combined with the other retrieved data, the collections were dispatched into smaller individual subsets, which were also separately released as parts of the Open Culture collection (Table 8). The Open Culture data in Common Corpus have been post-processed and filtered, as described below, which results in a slightly different final word and token count:

- **French PD.** This corpus is based on the training corpus for gallicagram[28]. It comprises 289,000 books from the French National Library (Gallica). This initial aggregation was made possible thanks to the open data program of the French National Library and the consolidation of public domain status for cultural heritage works in the EU following the 2019 Copyright Directive (Art. 14).

[17] https://publications.europa.eu/en
[18] http://eurovoc.europa.eu/
[19] https://op.europa.eu/en/web/cellar
[20] https://huggingface.co/datasets/EuropeanParliament/Eurovoc
[21] https://echanges.dila.gouv.fr/OPENDATA/
[22] https://huggingface.co/datasets/pile-of-law/pile-of-law
[23] https://digitallibrary.un.org/?ln=en
[24] https://archives.eui.eu/
[25] https://www.consilium.europa.eu/en/general-secretariat/corporate-policies/transparency/open-data/
[26] https://www.oecd.org/en/data/datasets.html?orderBy=mostRelevant&page=0
[27] https://api.bnf.fr
[28] https://shiny.ens-paris-saclay.fr/app/gallicagram

Table 8: Subsets of Open Culture with language coverage, type of document, and token count.

Corpus	Language	Domain	Tokens
English PD	English	Books and Newspapers	174.2B
US PD Books	English	Books	82.2B
French PD Books	French	Books	24.0B
French PD Newspapers	French	Newspapers	110.8B
French PD Diverse	French	Books and Newspapers	69.6B
LoC Books	English	Books	10.6B
US PD Newspapers	English	Newspapers	199.3B
New Zealand PD Newspapers	English, Māori	Newspapers	12.6B
Europeana Newspapers	Multilingual	Newspapers	21.0B
German PD Newspapers	German	Newspapers	18.4B
German PD	German	Books	58.0B
Portuguese PD	Portuguese	Books and Newspapers	2.6B
Spanish PD Newspapers	Spanish	Newspapers	8.0B
Spanish PD Books	Spanish	Books	15.4B
Italian PD	Italian	Books	18.2B
Dutch PD	Dutch	Books and Newspapers	2.7B
BnL Newspapers	German, French, Luxembourgish	Newspapers	0.3B
Danish PD	Danish	Books and Newspapers	0.5B
Serbian PD	Serbian	Books and Newspapers	0.3B
Czech PD	Czech	Books and Newspapers	0.7B
Greek PD	Greek	Books and Newspapers	4.2B
Multilingual PD	Multilingual	Books and Newspapers	8.4B
Polish PD	Polish	Books and Newspapers	5.9B
Latin PD	Latin	Books	27.2B
Russian PD	Russian	Books	1.9B
Arabic PD	Arabic	Books	0.3B

- **French PD Newspapers.** This dataset was also based on the Gallicagram corpus. It comprises nearly three million unique newspaper and periodical editions from the French National Library (Gallica).
- **LoC Books.** This dataset comprises 140,000 English books, digitized by the Library of Congress. The books come from the Selected Digitized Books Collection[29]. The dataset was curated by using the Library of Congress JSON API. This dataset contains only the books in the English collection. The dataset was compiled by Sebastian Majstorovic.
- **US PD Newspapers.** This dataset comprises 21 million digitized newspapers from Chronicling America[30]. The newspapers were digitized by the Library of Congress. The dataset can be fully explored through an original corpus map created by Nomic AI[31]. The dataset is mostly in English, but it also contains articles in other languages, mostly German and Spanish. The articles were published between the years 1690 and 1963.
- **New Zealand PD Newspapers.** This dataset comprises historic newspapers from New Zealand and the Pacific from the 19th and 20th centuries. The data were made available by the National Library of New Zealand as part of Papers Past[32]. The articles are primarily in English, but include some articles in te reo Māori.
- **Europeana Newspapers.** This dataset contains over 1,000 digitized newspapers from 23 libraries around Europe. It contains articles in at least 40 languages, and its articles were

[29] https://www.loc.gov/collections/selected-digitized-books/about-this-collection/

[30] https://chroniclingamerica.loc.gov/

[31] https://atlas.nomic.ai/data/aaron/pdnews-21286k-tr2k-addmeta/map

[32] https://paperspast.natlib.govt.nz/newspapers

published between 1618 and 1990 (Neudecker, 2016). The original sources are available via Europeana, and were made available by Big Science[33].

- **German PD Newspapers.** This dataset contains articles from 4,299,653 issues from over 1900 different newspapers. The articles come from the German Digital Library, hosted by Deutsches Zeitungsportal[34]. The articles were originally published between 1794 and 1957. This dataset was curated and first made available by Sebastian Majstorovic[35].
- **German PD.** This dataset contains texts from various sources, including the Mannheim Corpus of Historical Newspapers and Magazines[36] (Mannheimer Korpus Historischer Zeitungen und Zeitschriften). This dataset is made up of 21 German newspapers and magazines. The texts were originally published between 1737 and 1905. The corpus was originally digitized between 2009 and 2011. The corpus was made available by the Institut für Deutsche Sprache in 2013.
- **Spanish PD Books.** This dataset contains 302,640 individual texts from various sources, including the leading cultural heritage institution Biblioteca Digital Hispánica[37] (BDH). To ensure that these texts are in the public domain, we have retained exclusively titles published prior to 1884.
- **Dutch PD.** This dataset contains approximately 176,000 books and 540,000 periodicals, which come from various sources including Delpher[38]. Delpher is a repository of digitized printed material from the Netherlands, which is maintained by the Koninklijke Bibliotheek, the national library of the Netherlands. To ensure that these texts are in the public domain, we have retained exclusively titles published prior to 1884.
- **BnL Newspapers.** This dataset contains 630,709 articles from 21 different newspaper titles and 24,415 unique issues. The articles were digitized by the National Library of Luxembourg (BnL) as part of their Open Data Initiative[39]. OCR was done using Nautilus-OCR[40]. The articles are in German, French, and Luxembourgish. The newspapers were originally published between 1841 and 1879. The dataset was published and made accessible by BigScience.
- The rest of the datasets, including French PD Diverse, Portuguese PD, Italian PD, Polish PD, Danish PD, Swedish PD, Serbian PD, Czech PD, and Multilingual PD, come from various sources, including several European national libraries and cultural heritage institutions. To ensure that these texts are in the public domain, we have retained exclusively titles published prior to 1884.

E.3 Open Science

In Table 9, we present the total token counts per collection inside of the Open Science part of Common Corpus.

E.4 Open Code

Table 10 shows the number of tokens for the top ten coding languages and frameworks in Open Code.

F Open Culture Verification

Here, we describe the rights verification process that we applied for cultural data objects:

[33] https://huggingface.co/datasets/biglam/europeana_newspapers

[34] https://www.deutsche-digitale-bibliothek.de/newspaper

[35] https://huggingface.co/datasets/storytracer/German-PD-Newspapers

[36] https://repos.ids-mannheim.de/fedora/objects/clarin-ids:mkhz1.00000/datastreams/CMDI/content

[37] https://www.bne.es/fr/catalogues/biblioteca-digital-hispanica

[38] https://www.digitisednewspapers.net/histories/delpher/

[39] https://data.bnl.lu/

[40] https://github.com/natliblux/nautilusocr

Table 9: Token count by dataset Open Science.

Dataset	Tokens
OpenAlex	191,616,437,384
Open Science Pile	11,096,766,324
Open Science French	46,961,690,792
Open Science Spanish	16,523,491,767
Open Science German	7,806,446,050
ArXiv	7,188,731,472
Total	281,193,563,789

Table 10: Token counts by programming language or framework.

Language	Tokens
Java	35,697,451,454
JavaScript	28,894,772,110
Python	26,681,331,771
C++	25,481,950,314
C	23,277,000,113
PHP	23,077,121,733
C#	16,806,995,110
Go	11,200,587,099
Rust	3,888,428,173
Ruby	3,718,918,983

- **Author life + 70 years for all non-US authors.** Among most signatories of the Berne Convention for the Protection of Literary and Artistic Works[41], this is the most common approach to determining documents in the public domain. This approach requires not only identifying the author but also their date of death. On top of the information already made available by cultural heritage institutions, we also implemented an internal data reconciliation pipeline based on the complete dump of Wikidata.
- **All publications after 1884.** In cases where the author could not be identified or for collective works like newspapers, we applied a "universal" public domain rule based on 70 years prior to the current term of the author's life + 70 years. Simplified rules like these are commonly applied in cultural heritage projects, especially for the release of newspaper collections.
- **Publication + 95 years for US authors.** This is the copyright-based approach currently in place in the US. For an international project, this will only affect US-born authors. Due to a lack of further legal expertise, we did not attempt to include works whose copyright might not have been renewed.
- **No digitization rights.** Following on the 2019 Copyright Directive (Art. 14) and common practice among GLAM reusers like Wikimedia Commons, we consider that the simple act of digitization does not provide any additional rights.

G Cleaning and Curation

G.1 Text Segmentation

Here is an example input text for the Segmentext model:

> In this respect, the insurance business investment portfolio can be considered conservatively managed as it is largely composed of corporate, sovereign, and

[41] https://www.wipo.int/treaties/en/ip/berne/

> supranational bonds, term loans as well as demand deposits. Following the previous year, the group continued to diversify its holdings into investment-grade corporate bonds. It should be noted that bonds and term loans are held to maturity in accordance with the group's business model policy of "inflows".
>
> Technical liabilities on insurance contracts.
>
> The guarantees offered cover death, disability, redundancy, and unemployment as part of a loan protection insurance policy. These types of risk are controlled through the use of appropriate mortality tables, statistical checks on loss ratios for the population groups insured, and through the insurance program.
>
> Liability adequacy test.
>
> A goodness-of-fit test aimed at ensuring that insurance liabilities are adequate with respect to current statements of future cash flows generated by the insurance contracts is performed at each statement of account. Future cash flows resulting from the contracts take into account the guarantees and options inherent therein. In the event of inadequacy, the potential losses are fully recognized in the income statement. The modeling of future cash flows in the insurance liability adequacy test are based on the following assumptions: At the end of 2022, this liability adequacy test did not reveal any anomalies.
>
> Income statement.
>
> The income and expenses recognized for the insurance contracts issued by the group appear in the income statement in "Net income of other activities" and "Net expense of other activities".
>
> Risk management.
>
> The group adopts a "prudent approach" to its management of the risks to which it could be exposed through its insurance activities. Risk of counterparty. As stated above, insurance companies only invest in assets (bank deposits, sovereign bonds, supranational agencies, or corporate bonds).

Example output:

Editorial Segmentation

[Text] In this respect, the insurance business investment portfolio can be considered conservatively managed as it is largely composed of corporate, sovereign, and supranational bonds, term loans as well as demand deposits. Following the previous year, the group continued to diversify its holdings into investment-grade corporate bonds. It should be noted that bonds and term loans are held to maturity in accordance with the group's business model policy of "inflows".

[Title] **Technical liabilities on insurance contracts.**

[Text] The guarantees offered cover death, disability, redundancy, and unemployment as part of a loan protection insurance policy. These types of risk are controlled through the use of appropriate mortality tables, statistical checks on loss ratios for the population groups insured, and through the insurance program.

[Title] **Liability adequacy test.**

[Text] A goodness-of-fit test aimed at ensuring that insurance liabilities are adequate with respect to current statements of future cash flows generated by the insurance contracts is performed at each statement of account. Future cash flows resulting from the contracts take into account the guarantees and options inherent therein. In the event of inadequacy, the potential losses are fully recognized in the income statement. The modeling of future cash flows in the insurance liability adequacy test are based on the following assumptions: At the end of 2022, this liability adequacy test did not reveal any anomalies.

[Title] **Income statement.**

[Text] The income and expenses recognized for the insurance contracts issued by the group appear in the income statement in "Net income of other activities" and "Net expense of other activities".

[Title] **Risk management.**

[Text] The group adopts a "prudent approach" to its management of the risks to which it could be exposed through its insurance activities.

[Title] **Risk of counterparty.**

[Text] As stated above, insurance companies only invest in assets (bank deposits, sovereign bonds, supranational agencies, or corporate bonds).

G.2 OCR Error Detection

OCRoscope. To illustrate this approach, this long text is correctly identified as French with ¿99% confidence by `cld2`, as despite the many mistakes, there are enough non-ambiguous French words:

> NOUVELLES POLI TI QÛ E S. Suede. Stockholm , le 2 5 décembre 1792. Le général Toll ira à Varsovie en quarté d'envoyé de la Suede auprès du roi et de la république ; A 1 même rey.u l'ordre de s'y rendra incessamment. 11 paraît que k Uc-régeik a des craintes ; il a fait venir chez lji les membres c Ij"‘ tribunal 4e la cour , et leur a rtmis son lesfca n at. La fermentation qu'a causée 1 ,'ari r?tavh n k M p v riote Thorild tî'est pas appaisée y le luigage qv'il a yailé an duc-régent a été bien entendu par le peu) k y ir M (U i n'entendrait pas l'apostrdphe suivante ? ttRxc3xa7nd ¿la libuk à r otre raison , et ne et nous force pas de i'ache'ef r i te n :e sang,.
>
> Le duc a fait x,épa4idre sur-le-champ une fjtbprijuun à te us les habitans di$ Toyaume , pour les detourntr de mr laisser sé luire par de fa,ux bruits et des jugemens pe rver$, e i en même temps l'ordre a. été donné à la garnison de charger et de se tenir prête à marcher.
>
> (Mercure Français, 1793, January 25th)

Yet one short n-gram ("n k M p v riote Thorild") is classified as unknown by cld2.

OCRerrcr. The following is a low-error example sentence taken from Common Corpus:

> They did not approach cer, but turned away and passed irom her presence, filled with sorrow and moved with sympathy, which her intense emotions seemed to communicate to even these thoughtless young men of the tho plains.

And the OCRerrcr detection (with formatting for clarity):

> They did not approach <er>cer,</er> but turned away and passed <er>irom</er> her presence, filled with sorrow and moved with sympathy, which her intense emotions seemed to communicate to even these thoughtless young men of the <er>tho</er> plains.

G.3 OCR Correction

Here is an example of text containing various OCR errors:

> Theguaran tees offered cover death,disability,r e dundancy andunem ployment aspartof aloanprotect ion insurance policy. These types o f risk are controlled throu ghthe use o f app ropriate morta litytables,statistica lchecksonloss rat ios for thepopulation groups insure dandthrough ar e insurance program.

And here is the text corrected by our model, OCRonos:

> The guarantees offered cover death, disability, redundancy, and unemployment as part of a loan protection insurance policy. These types of risk are controlled through the use of appropriate mortality tables, statistical checks on loss ratios for the population groups insured, and through the insurance program.

H Evaluations

In Tables 11, 12, 13, and 14, we present per-language scores for the studied benchmarks. On MultiBLIMP, most of the scores are significantly above random (0.5); therefore, we also highlight the best and second-best scores.

Table 11: Multilingual benchmarking results on MultiBLIMP (ISO 639 language codes a*–i* in alphabetical order). "Ours" refers to our models pre-trained on Common Corpus. Within each model group, the best score is in **bold**, and the second-best is underlined.

Model	**Ours**	**Gemma 3**	**XGLM**	**BLOOM**	**Ours**	**Gemma 3**	**XGLM**	**OLMo**
Parameters	350M	270M	564M	560M	1.2B	1B	1.7B	1B
abk	0.550	**0.750**	0.475	0.525	0.675	0.675	0.325	**0.800**
aln	0.733	**0.755**	0.709	0.700	0.728	**0.750**	0.675	0.690
amh	0.946	0.929	0.911	**0.973**	**1.000**	0.955	0.848	0.964
apu	**0.964**	**0.964**	0.893	0.786	**0.964**	0.929	**0.964**	0.893
aqz	0.214	0.357	0.429	**0.500**	0.429	0.429	**0.714**	0.214
arb	0.877	0.913	0.895	**0.923**	0.900	**0.951**	0.887	0.782
azz	0.734	**0.744**	0.729	0.720	0.729	0.758	**0.773**	0.686
bel	**0.799**	0.795	0.574	0.608	0.853	**0.896**	0.577	0.611
ben	0.571	0.762	**1.000**	0.810	0.762	**0.905**	0.857	0.524
bho	**0.676**	0.647	0.588	0.588	0.706	**0.794**	0.618	0.588
bor	**0.722**	0.631	0.697	0.610	**0.697**	0.627	0.680	0.668
bre	**0.942**	0.815	0.554	0.604	0.938	**0.946**	0.615	0.685
bua	0.680	**0.718**	0.670	**0.718**	0.670	0.641	**0.699**	0.660
bul	0.872	0.880	**0.969**	0.623	0.897	0.945	**0.976**	0.735
cat	0.885	0.852	**0.961**	0.950	0.919	0.931	**0.953**	0.735
ces	**0.824**	0.808	0.579	0.597	0.858	**0.891**	0.603	0.668
chu	**0.670**	0.648	0.582	0.635	0.659	**0.663**	0.593	0.632
cym	**0.771**	0.730	0.633	0.611	**0.828**	0.796	0.610	0.796
dan	0.980	**1.000**	0.840	0.800	0.980	**1.000**	0.740	0.940
deu	**0.967**	0.949	0.961	0.754	0.977	**0.981**	0.969	0.886
egy	0.409	0.409	0.409	**0.455**	**0.500**	0.455	0.409	0.455
ell	0.931	0.937	**0.985**	0.676	0.948	0.975	**0.984**	0.842
eng	**0.981**	0.979	0.973	0.960	0.983	**0.987**	0.974	0.984
est	0.729	0.699	**0.885**	0.561	0.800	0.800	**0.915**	0.587
eus	0.916	0.927	**0.963**	0.952	0.916	0.938	**0.982**	0.905
fao	**0.707**	0.647	0.509	0.556	0.772	**0.806**	0.552	0.681
fas	0.756	**0.810**	0.567	0.577	0.837	**0.919**	0.565	0.655
fin	0.736	0.744	**0.947**	0.562	0.809	0.893	**0.935**	0.645
fra	**0.994**	0.963	0.963	0.984	**0.993**	0.989	0.976	0.928
frm	**0.997**	0.741	0.745	0.905	**0.997**	0.847	0.820	0.765
fro	**0.782**	0.701	0.686	0.709	**0.822**	0.725	0.694	0.679
gla	**0.955**	0.924	0.924	0.939	0.909	0.924	0.939	**0.970**
gle	0.750	**0.821**	0.750	0.786	0.679	**0.750**	**0.750**	0.714
glg	**0.849**	0.807	0.798	0.788	0.879	**0.895**	0.789	0.754
got	0.630	**0.642**	0.588	0.599	**0.645**	0.631	0.580	0.598
grc	**0.824**	0.719	0.683	0.623	**0.887**	0.758	0.711	0.707
guj	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.857	**1.000**	**1.000**
hbo	0.711	**0.712**	0.683	0.706	0.737	**0.764**	0.629	0.679
hbs	**0.892**	0.848	0.619	0.603	0.922	**0.929**	0.616	0.723
heb	**0.853**	0.829	0.609	0.642	**0.876**	0.868	0.585	0.667
hin	0.854	0.934	**0.975**	0.971	0.916	0.966	**0.977**	0.748
hit	0.620	**0.640**	0.520	0.600	**0.560**	0.480	0.540	0.540
hsb	**0.683**	0.677	0.624	0.629	0.667	**0.694**	0.624	0.591
hun	**0.928**	0.867	0.728	0.692	**0.938**	0.925	0.686	0.740
hye	0.883	**0.898**	0.631	0.623	0.929	**0.946**	0.662	0.671
hyw	**0.813**	0.781	0.583	0.608	**0.894**	0.859	0.551	0.629
isl	0.710	**0.751**	0.653	0.660	0.767	**0.863**	0.636	0.667
ita	**0.925**	0.910	0.915	0.670	0.952	**0.965**	0.915	0.791

Table 12: Multilingual benchmarking results on MultiBLIMP (ISO 639 language codes k*–y* in alphabetical order). "Ours" refers to our models pre-trained on Common Corpus. Within each model group, the best score is in **bold**, and the second-best is underlined.

Model	**Ours**	**Gemma 3**	**XGLM**	**BLOOM**	**Ours**	**Gemma 3**	**XGLM**	**OLMo**
Parameters	350M	270M	564M	560M	1.2B	1B	1.7B	1B
kat	0.931	**0.951**	0.917	0.760	**0.951**	**0.951**	0.809	0.907
kaz	0.705	**0.792**	0.647	0.682	0.780	**0.844**	0.682	0.688
kir	**0.930**	0.843	0.914	0.919	0.935	0.924	0.843	**0.941**
kmr	**0.710**	0.662	0.588	0.579	**0.761**	0.748	0.577	0.588
koi	**0.628**	0.488	0.605	0.558	0.558	**0.651**	0.628	0.605
kpv	**0.641**	0.591	0.553	0.547	**0.700**	0.628	0.581	0.547
krl	0.650	0.612	**0.688**	0.573	0.612	0.642	**0.704**	0.573
kxh	**0.483**	0.433	0.475	0.333	0.458	0.450	0.442	**0.483**
lat	**0.874**	0.651	0.578	0.575	**0.925**	0.730	0.568	0.625
lav	**0.791**	0.747	0.616	0.604	0.844	**0.862**	0.611	0.623
lij	**0.783**	0.744	0.669	0.638	0.780	**0.807**	0.701	0.665
lit	**0.928**	0.848	0.745	0.740	**0.947**	0.932	0.736	0.779
mar	**0.737**	0.717	0.735	0.667	0.713	**0.776**	0.726	**0.776**
mdf	0.537	**0.622**	0.524	0.537	**0.622**	0.585	0.561	0.500
mkd	0.923	**0.974**	0.769	0.769	0.821	**1.000**	0.590	0.718
myv	0.608	**0.614**	0.565	0.560	**0.636**	0.619	0.532	0.547
nds	**0.736**	0.729	0.674	0.663	**0.749**	0.732	0.674	0.700
nhi	0.526	0.579	0.474	**0.632**	0.553	**0.579**	0.447	0.500
nld	**0.924**	0.912	0.620	0.627	0.954	**0.963**	0.663	0.829
olo	0.679	0.668	**0.795**	0.611	0.753	0.711	**0.842**	0.595
orv	**0.733**	0.721	0.690	0.636	**0.757**	0.744	0.707	0.667
ota	0.879	**0.929**	0.899	0.848	0.939	**0.949**	0.889	0.828
pcm	**1.000**	**1.000**	**1.000**	0.923	**1.000**	0.962	**1.000**	0.885
pol	**0.892**	0.849	0.624	0.634	0.930	**0.931**	0.628	0.725
por	0.948	0.933	0.939	**0.955**	0.965	**0.972**	0.920	0.872
quc	**0.779**	0.672	0.649	0.740	**0.740**	0.664	0.679	0.656
ron	0.868	**0.874**	0.638	0.608	0.903	**0.928**	0.640	0.793
rus	0.921	0.916	**0.937**	0.727	0.952	**0.963**	0.954	0.819
sah	0.688	0.771	0.736	**0.792**	0.708	0.681	0.701	**0.764**
san	0.657	**0.666**	0.612	0.609	0.670	**0.678**	0.618	0.620
slk	**0.797**	0.739	0.528	0.570	0.824	**0.861**	0.533	0.588
slv	**0.882**	0.796	0.618	0.636	**0.903**	0.854	0.622	0.711
sme	0.689	**0.705**	0.653	0.660	0.681	**0.700**	0.668	0.659
sms	**0.833**	0.802	0.779	0.757	**0.821**	0.779	0.764	0.768
spa	0.959	0.945	0.950	**0.966**	0.970	**0.973**	0.956	0.896
sqi	0.786	**0.823**	0.494	0.588	0.823	**0.881**	0.539	0.765
swe	**0.995**	**0.995**	0.970	0.950	0.995	**1.000**	0.985	0.990
tam	0.942	0.942	**0.969**	0.966	0.932	0.953	**0.976**	0.746
tpn	**0.111**	0.000	**0.111**	**0.111**	**0.222**	0.111	0.000	0.000
ttc	0.478	0.478	**0.493**	0.449	0.449	0.464	0.435	**0.478**
tur	0.766	0.804	**0.829**	0.700	0.814	**0.908**	0.857	0.716
uig	**0.764**	0.760	0.722	0.732	0.755	**0.757**	0.686	0.740
ukr	0.874	**0.892**	0.640	0.606	0.911	**0.946**	0.648	0.704
urb	**0.538**	0.462	0.462	0.231	**0.538**	0.462	0.462	0.462
urd	0.858	0.925	**0.956**	0.935	0.896	**0.964**	0.960	0.736
uzb	**0.900**	0.880	0.780	0.720	0.900	**0.940**	0.760	0.880
vep	0.572	**0.588**	**0.588**	0.481	**0.631**	0.626	0.567	0.524
wbp	**0.250**	0.000	0.167	**0.250**	**0.250**	0.083	**0.250**	**0.250**
wol	**0.892**	0.881	0.851	0.891	0.868	**0.879**	0.854	0.823
xcl	**0.679**	0.674	0.585	0.636	**0.711**	0.702	0.613	0.616
xnr	**0.791**	0.756	0.779	0.616	0.744	**0.814**	0.721	0.733
xpg	0.820	**0.900**	0.820	0.860	0.880	**0.900**	**0.900**	**0.900**
yrl	0.689	**0.722**	0.594	0.604	**0.683**	0.669	0.626	0.601

Table 13: Multilingual benchmarking results on XStoryCloze. "Ours" refers to our models pre-trained on Common Corpus. Languages are represented as two-letter codes in ISO 639.

Model	**Ours**	**Gemma 3**	**XGLM**	**BLOOM**	**Ours**	**Gemma 3**	**XGLM**	**OLMo**
Parameters	350M	270M	564M	560M	1.2B	1B	1.7B	1B
ar	0.475	0.492	0.500	0.521	0.477	0.572	0.525	0.473
ca	0.514	0.513	0.567	0.561	0.535	0.600	0.602	0.509
en	0.569	0.614	0.606	0.612	0.617	0.698	0.645	0.704
es	0.520	0.558	0.549	0.555	0.543	0.628	0.593	0.556
eu	0.516	0.524	0.531	0.538	0.514	0.531	0.561	0.503
gl	0.490	0.486	0.461	0.467	0.532	0.576	0.484	0.459
hi	0.509	0.541	0.520	0.549	0.515	0.598	0.557	0.493
id	0.500	0.544	0.542	0.555	0.518	0.631	0.583	0.498
my	0.492	0.507	0.515	0.475	0.498	0.518	0.537	0.475
ru	0.503	0.547	0.562	0.488	0.531	0.634	0.600	0.512
sw	0.500	0.507	0.531	0.500	0.510	0.551	0.563	0.494
te	0.542	0.562	0.559	0.557	0.553	0.592	0.581	0.532
zh	0.491	0.536	0.533	0.545	0.494	0.594	0.561	0.511

Table 14: Multilingual benchmarking results on XCopa. "Ours" refers to our models pre-trained on Common Corpus. Languages are represented as two-letter codes in ISO 639.

Model	**Ours**	**Gemma 3**	**XGLM**	**BLOOM**	**Ours**	**Gemma 3**	**XGLM**	**OLMo**
Parameters	350M	270M	564M	560M	1.2B	1B	1.7B	1B
es	0.566	0.590	0.604	0.620	0.614	0.678	0.664	0.524
et	0.518	0.498	0.554	0.488	0.500	0.536	0.568	0.480
eu	0.504	0.514	0.512	0.502	0.518	0.502	0.534	0.516
ht	0.522	0.504	0.548	0.500	0.524	0.518	0.556	0.534
id	0.534	0.578	0.574	0.596	0.558	0.690	0.646	0.544
it	0.542	0.524	0.536	0.502	0.562	0.648	0.536	0.488
qu	0.522	0.492	0.492	0.500	0.500	0.502	0.522	0.506
sw	0.528	0.546	0.530	0.516	0.538	0.544	0.562	0.510
ta	0.536	0.560	0.562	0.558	0.556	0.566	0.550	0.550
th	0.546	0.538	0.550	0.538	0.550	0.584	0.580	0.532
tr	0.530	0.566	0.544	0.528	0.542	0.606	0.536	0.530
vi	0.550	0.598	0.584	0.602	0.526	0.694	0.630	0.494
zh	0.536	0.564	0.554	0.588	0.544	0.640	0.584	0.522