

---

# Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training

---

Pierre-Carl Langlais

Carlos Rosas Hinostroza

Mattia Nee

Catherine Arnett Pavel Chizhov Eliot Krzysztof Jones Irène Girard

David Mach

Anastasia Stasenko

Ivan P. Yamshchikov

PleIAS, Paris, France <https://pleias.fr/>

## Abstract

Large Language Models (LLMs) are pre-trained on large amounts of 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<sup>1</sup>, the largest open dataset for language model 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 main European languages to low-resource ones rarely present 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 technical report, we present the detailed provenance of data assembling and the details of dataset filtering and curation. Being already used by such industry leaders as Anthropic and multiple LLM training projects, we believe that Common Corpus will become a critical infrastructure for open science research in LLMs.

## 1 Introduction

Large Language Models have been defined by large 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 and 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 (e.g. Books3). Until 2025, LLM training data has grown on a logarithmic curve. The latest generation of publicly documented language models including DeepSeek v3 (DeepSeek-AI, 2025), Gemma 3 (Google, 2025), Llama 4 (Meta AI, 2025) or Qwen 3 (Qwen, 2025) have been trained on 14-36 trillion tokens. Even the gaining popularity sub-category of *small*

---

<sup>1</sup>[https://huggingface.co/datasets/PleIAS/common\\_corpus](https://huggingface.co/datasets/PleIAS/common_corpus)

*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, about 3,000 times more data than recommended by the original Chinchilla laws (Hoffmann et al., 2022).

As data curation also became a primary concern, the collection, maintenance, processing, and filtering of data became one of the primary cost of standard language model training. This does not factor in even larger hidden costs: the 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 do 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 Open AI for copyright infringement, alleging that Open AI trained their models on their 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 language model 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.

These 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) and by multiple AI labs, including Meta and Microsoft, faced legal challenges (Brittain, 2023), and the original dataset was ultimately removed (Van der Sar, 2023). The LAION dataset was demonstrated to contain CSAM (Birhane et al., 2021; Thiel, 2023), and then was taken down (LAION, 2023). Later, it was re-released once suspected CSAM was removed (LAION, 2024). The Dutch-focused model GEITje was taken down (Rijgersberg, 2025) due to complaints about it having been trained 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 becoming increasingly more difficult 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 with clear licensing and in the English language. 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 not requiring attribution (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, an ongoing data collection of 1 trillion tokens from a variety of recent sources, including a filtered version of Common Crawl data (Creative Commons Common Crawl) (Kandpal and Raffel, 2025). These projects are all monolingual, restricting in effect the reach of open-data language models to the English-speaking audience. In contrast, the most ambitious multilingual collection of permissive content pre-dates Large Language

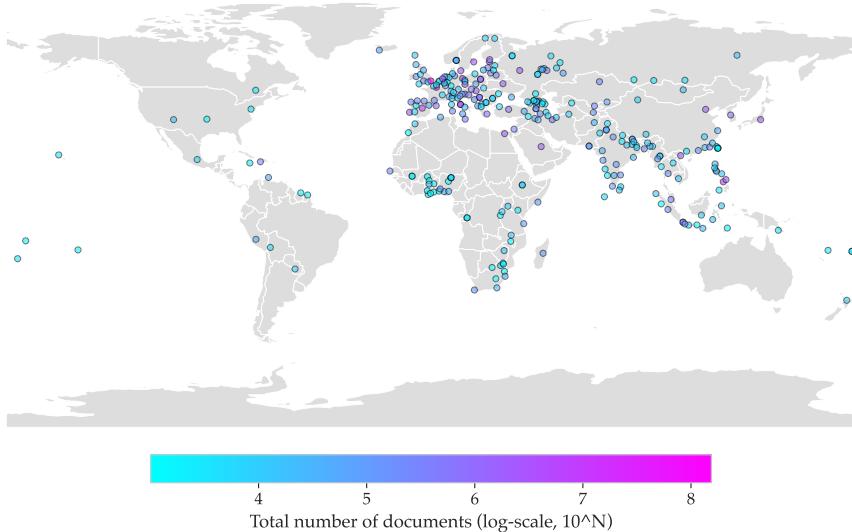


Figure 1: A schematic world map of languages in Common Corpus with a log-scaled distribution of document counts. For each language, we chose a city that is located in the region where this language is most specific to. To avoid outliers, we show only languages with 1000+ documents.

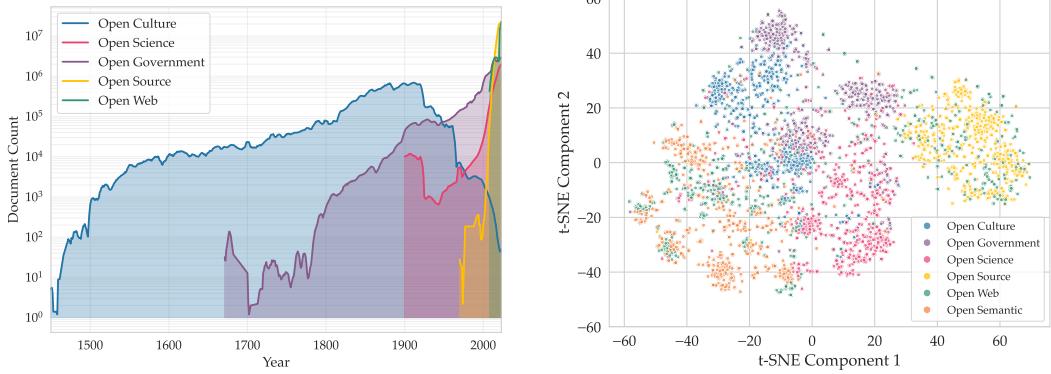
Models: C4C (2016), containing 12 million web pages in more than 50 languages filtered by Creative Commons Licenses ([Habernal et al., 2016](#)).

In this paper, we present **Common Corpus**, which has been an ongoing effort since 2024 (see Appendix A for the development history). It has grown to become the largest open dataset for pre-training Large Language Models at about **2 trillion tokens** and the only one in its size range available in multiple languages. Thus, Common Corpus is a significant contribution to the AI Commons. Through the release of this dataset, we show that open LLM research and development is possible while meeting legal and regulatory requirements—even in the EU, where AI regulations are currently the strictest. Here, we detail the composition of Common Corpus (Section 3), as well as the data provenance, license verification, and further processing (Section 4). Next, we describe the early impact of Common Corpus on the AI community (Section 5) and highlight the limitations of the work that has been done at the time of publication. Despite its size, Common Corpus is still far from covering the entire range of available resources for pre-training commons: 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 sources currently used for pre-training. By describing the unique challenges coming with the aggregation of large open source, we aim to inspire further responsible data collection initiatives.

## 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 processing, and important information about the contents of each dataset. Following the Open Source Initiative’s definition of open-source AI, our data is open in terms of openness of use, meaning 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 relevant 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



(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.

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
Open Semantic	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

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

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 and Berkouk, 2024](#); [Baack et al., 2025](#)). As PleIAs was one of the contributors at the Dataset Convening, 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. This is crucial, as 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 4.4).

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<sup>2</sup> 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.

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 2: Token counts for the ten most represented languages 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

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

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 needed for a particular use case.

Common Corpus is multilingual (see Figure 1 and Table 2 for language coverage). Unlike many pretraining mixes, which are composed mostly of English and code data, our dataset has strong representation for several languages, with the top nine languages constitute at least

<sup>2</sup>We report token counts in terms of the Pleias base tokenizer <https://huggingface.co/PleIAS/Pleias-350m-Preview> trained on a subsample of Common Corpus.

10B tokens each<sup>3</sup>. While we highlighted some of the issues faced in making open datasets above, all of these issues are exacerbated 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.

### 3 Provenance

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

#### 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 B.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, from the mid-20th century 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 a wide variety of other languages, including Arabic, Bengali, Latin, Persian, Russian, Sanskrit, and Urdu.

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

---

<sup>3</sup>Language distribution was computed using the fastText language identification model.

irrecoverable deterioration of the original support. For the detailed dataset composition, refer to Appendix B.2.

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 C).

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 and 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 on 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 for training data curation, due to its impact on performance on reasoning and advanced world knowledge tasks. 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 dumps made available by 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 (Workshop et al., 2023).

The Open Science collection was made possible largely due to the recent development of OpenAlex<sup>4</sup>, 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 and some European-specific resources not currently well indexed on OpenAlex (the exact distribution of token counts can be found in Appendix B.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.

---

<sup>4</sup><https://openalex.org/>

### 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 B.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.** YouTube Commons is our collection of audio transcripts of 2,063,066 videos uploaded on YouTube under a standardized CC-By license<sup>5</sup>.

**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 will be the integration of web archives filtered by permissive license. 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; the first multilingual C4 derivative)<sup>6</sup> All theses 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 different CC-license type or a blog home page with many articles and each article is licensed under 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.

---

<sup>5</sup><https://huggingface.co/datasets/PleIAs/YouTube-Commons>

<sup>6</sup><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

### 4.1 Text Segmentation

For text segmentation, we developed **Segmentext**<sup>7</sup> (see example in Appendix D.1). It was 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.

### 4.2 OCR Error Detection

We developed two pipelines to determine the digitization quality of different datasets in order to determine the appropriate treatment before datasets can be used for pre-training:

**OCRoscope**<sup>8</sup>. This is a tool based on the language identification tool cld2<sup>9</sup>. As cld2 cannot attribute non-recognized strings to any language, OCRoscope works by measuring the number of non-recognized 7-grams. OCRoscope provides a standardized rate of OCR quality and rate of non-character content for at least 80 languages. An example of OCPoscope work is presented in Appendix D.2. Complete processing by OCRoscope yielded a rate of 41% non-recognized 7-grams, which results in an OCR quality rate of 59%. In comparison, the self-estimated rate of OCR valid words by the French National Library is significantly higher (85%) for the whole document.

**OCRerrcr**<sup>10</sup>. This is a DeBERTa-v2-style language model with 400M parameters. OCRerrcr achieves higher accuracy than OCRoscope, but is more computationally intensive.

### 4.3 OCR Correction

We developed **OCRonos**<sup>11</sup> model based on Llama 3 8B (Grattafiori et al., 2024). OCRonos is versatile and supports the correction of OCR errors, wrong word cuts or merges, 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

<sup>7</sup><https://huggingface.co/PleIAs/Segmentext>

<sup>8</sup><https://github.com/Pleias/OCRoscope>

<sup>9</sup><https://pypi.org/project/pycld2/>

<sup>10</sup><https://huggingface.co/PleIAs/OCRerrcr>

<sup>11</sup><https://huggingface.co/PleIAs/OCRonos>

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 D.3. OCROnos contributes to making challenging resources usable for LLM applications and, more broadly, search retrieval. It is especially fitting in situations where the original PDF sources are too damaged for correct OCR or are not possible to retrieve.

OCRonos is generally faithful to the original material, provides sensible restitution of deteriorated text, and will rarely rewrite correct words. 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.

#### 4.4 PII Removal

Personally Identifiable Information (PII), which is 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 an analysis of 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<sup>12</sup>, an open-source state-of-the-art tool. 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.

#### 4.5 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 and Viviano, 2021)) and often suffers from low-quality and machine-generated text (Dodge et al., 2021). Public Domain data, *e.g.*, 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. Cultural norms surrounding the discussion of certain ethnic groups, women, and themes such as violence have changed dramatically. Many of these texts, therefore, do not meet modern ethical standards. Training 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 and stylistically diverse data for high model performance. We created a multilingual toxicity classifier, Celadon<sup>13</sup>, a DeBERTa-v3-small model (~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. Celadon and the training dataset<sup>14</sup> were released as parts of a separate work (Arnett et al., 2024).

---

<sup>12</sup><https://microsoft.github.io/presidio/>

<sup>13</sup><https://huggingface.co/PleIAs/celadon>

<sup>14</sup><https://huggingface.co/datasets/PleIAs/ToxicCommons>

## 5 Impact

As early as August 2024, the first version of Common Corpus was highlighted as a “massive” part of pre-training data commons in European languages, especially contributing to making French a high-resource language (Ali and Pysalo, 2024). In February 2025, Common Corpus was acknowledged as one of the main deliverables of the Paris AI Summit, as the “largest open database to train Large Language Models”<sup>15</sup>.

Common Corpus has already been incorporated in a wide range of LLM projects. As of May 2025, seven European SLMs and LLMs have been trained on at least some part of Common Corpus, and more will come throughout the year 2025. Among these are the models in the Pleias 1.0 family, including three SLMs (350 million, 1.2 billion, and 3 billion parameters), which were exclusively trained on Common Corpus. As such, they are fully reproducible in Europe. They were trained on the Jean Zay supercomputer in France (using H100 GPUs, under a Grand Challenge compute grant), and, specifically for the 1.2B model on the TractoAI’s cloud platform<sup>16</sup>. They come with a new tokenizer with 65,536 unique tokens, also trained on a representative sample of the Common Corpus for increased training efficiency and better multilingual support. During training, we experimented with different data mixtures of Common Corpus and derived a “refined” version for late training of about 1 trillion tokens, with more intensive, curation-based filtering. Two models derived from Pleias 1.0, Pleias-RAG-350M and Pleias-RAG-1B, are currently leading evaluations in their size range for retrieval augmented evaluation. Part of the performance can be attributed to the solid coverage of the main European language in Common Corpus, as well as a very extensive proportion of PDF-extracted sources that are very close to content commonly processed in the RAG production setting.

In January 2025, Barcelona Supercomputing Center released the Salamandra series of models under three different sizes: 2, 7, and 40 billion parameters (Gonzalez-Agirre et al., 2025). A major objective was enhanced support for a wide range of European languages, including lower-resource languages. The French part of *Open Culture*, French-PD, was incorporated into the overall data mixture. It was the main French dataset and the second-largest curated source. A French 7 billion parameters SLM, Lucie, incorporated a wider range of French sources from Common Corpus, not only the cultural heritage part but also some of the administrative sources (Gouvert et al., 2025). In addition, NeKo, a generalist language correction model trained by Nvidia, used the OCR-corrected part of Common Corpus both for training and evaluation (Lin et al., 2024). Anthropic also used Common Corpus in their experiments for feature visualization (Ameisen et al., 2025).

Common Corpus is also being used to create new datasets. For instance, YouTube Commons was used to create two major multimodal datasets: FineVideo (43000 YouTube videos)<sup>17</sup> and Mosel (1 million audio hours, half from YouTube Commons) (Gaido et al., 2024), which was used to train the FAMA series of speech foundational models (Papi et al., 2025).

## 6 Conclusion

Common Corpus is a significant contribution to the open LLM research community. Since the release of the first version, the data from Common Corpus has been used in numerous LLM projects by several European labs and industry leaders like Anthropic. 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. 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.

---

<sup>15</sup><https://www.elysee.fr/admin/upload/default/0001/17/373d6bcd8ea7f84a701c050bb9cabccf3ed95c2b.pdf>

<sup>16</sup><https://tracto.ai/>

<sup>17</sup><https://huggingface.co/datasets/HuggingFaceFV/finevideo>

## Limitations

Common Corpus is far from collecting the whole range of available open data, which we call the *open data paradox*. Continuing the work, we are currently collecting more data in European languages as part of the LLMs4EU project<sup>18</sup>. In addition to this, even though the language diversity of the data is large, the coverage for low-resource languages is still shallow. We are developing in this direction, planning for enlargement of the corpora in African languages (*e.g.*, Swahili, Wolof, Bambara). Furthermore, the amount of collected data (2 trillion tokens) is suitable for pre-training of models of limited size, as in the Pleias model family (see Section 5). Larger models, however will require significantly larger amounts of data. Therefore, future collection of permissible data is highly encouraged by this work.

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, as we did for the RAG models (see Section 5).

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 and PII removal), 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).

## Acknowledgments and Disclosure of Funding

Common Corpus is part of the Current AI initiative to accelerate the development of the building blocks of open AI — notably data — that serves and is shaped by the public interest. It was built up with the support and concerted efforts of AI Alliance, the state start-up LANGU:IA (start-up d'Etat), supported by the French Ministry of Culture and DINUM, as part of the prefiguration of the service offering of the Alliance for Language technologies EDIC (ALT-EDIC). The creation of the OpenCulture datasets was made possible by the support of LANGU:IA. The Wikidata and Wikipedia datasets were made in partnership with Wikimedia Enterprise and Wikimedia Germany. The corpus was stored and processed with the generous support of Genci (Jean-Zay, Idris, Eviden), Scaleway, and Tracto AI. The collection of the corpus has been largely facilitated thanks to major organizations committed to an open science approach for AI, namely AI Alliance, Mozilla, HuggingFace, Occiglot, and Eleuther AI.

## 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, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 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/>.

Wazir Ali and Sampo Pyysalo. A Survey of Large Language Models for European Languages, August 2024. URL <http://arxiv.org/abs/2408.15040>. arXiv:2408.15040 [cs].

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus,

---

<sup>18</sup><https://www.openaire.eu/llms4eu>

Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. <https://transformer-circuits.pub/2025/attribution-graphs/methods.html>.

Catherine Arnett, Eliot Jones, Ivan P Yamshchikov, and Pierre-Carl Langlais. Toxicity of the commons: Curating open-source pre-training data. *arXiv preprint arXiv:2410.22587*, 2024.

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, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/biderman23a.html>.

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, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1636. URL <https://aclanthology.org/P19-1636>.

Team DeepSeek-AI. DeepSeek-V3 Technical Report, February 2025. URL <http://arxiv.org/abs/2412.19437>. arXiv:2412.19437 [cs].

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, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 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>.

Marco Gaido, Sara Papi, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih, and Matteo Negri. MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages, October 2024. URL <http://arxiv.org/abs/2410.01036>. arXiv:2410.01036 [cs].

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.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. Salamandra Technical Report, February 2025. URL <http://arxiv.org/abs/2502.08489>. arXiv:2502.08489 [cs].

Gemma Team Google. Gemma 3 Technical Report, March 2025. URL <http://arxiv.org/abs/2503.19786>. arXiv:2503.19786 [cs].

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and OpenLLM-France community. The Lucie-7B LLM and the Lucie Training Dataset: Open resources for multilingual language generation, March 2025. URL <http://arxiv.org/abs/2503.12294>. arXiv:2503.12294 [cs].

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 Srivankumar, 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, editors, *Proceedings of the Tenth International*

*Conference on Language Resources and Evaluation (LREC'16)*, pages 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, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 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>.

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*, pages 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, Orevaghene 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/>.

Yen-Ting Lin, Chao-Han Huck Yang, Zhehuai Chen, Piotr Zelasko, Xuesong Yang, Zih-Ching Chen, Krishna C. Puvvada, Szu-Wei Fu, Ke Hu, Jun Wei Chiu, Jagadeesh Balam, Boris Ginsburg, and Yu-Chiang Frank Wang. NeKo: Toward Post Recognition Generative Correction Large Language Models with Task-Oriented Experts, November 2024. URL <http://arxiv.org/abs/2411.05945>. arXiv:2411.05945 [cs].

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 Tetraault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 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/](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, editors, *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 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, Sébastien 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, editors, *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)*, pages 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>.

Team Meta AI. 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, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 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.

Sara Papi, Marco Gaido, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih, and Matteo Negri. Fama: The first large-scale open-science speech foundation model for english and italian, 2025. URL <https://arxiv.org/abs/2505.22759>.

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=2sghJ1yY0r>.

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-times-about-face/>.

Team Qwen. Qwen3: Think Deeper, Act Faster, April 2025. URL <https://qwenlm.github.io/blog/qwen3/>. Section: blog.

- 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>.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Beglin, 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>.
- BigScience Workshop, :, 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 Abdulkummin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, et al. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL <https://arxiv.org/abs/2211.05100>.
- 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=B18u7ZRlbM>.

## A Development

Common Corpus was first conceptualized in May 2023 as a French corpus for the French public services. The initial aim was to create a dataset of 100-140B tokens and use that data to train a reproducible small language model, modeled on the Pythia suite ([Biderman et al., 2023](#)). The initial nucleus of Common Corpus was French-PD a large collection of 85B words from monographs and newspapers in the public domain<sup>19</sup>.

A collection of various open datasets was first published under the name *Common Corpus* in March 2024<sup>20</sup>. At the time, it was a multilingual expansion of *French-PD*, retaining the initial focus on cultural heritage data but containing now about 500B words in the public domain from a variety of open collections in English, German, French, Italian, Spanish and other European languages. Throughout 2024, additional official releases gradually expanded this initial pretraining commons beyond the public domain, including Youtube Commons (in April 2024) and Finance Commons (in July 2024).

In November 2024, *Common Corpus* became an unified corpus, through the convergence of newly collected datasets by PleIAs and pre-existing ones (especially Wikimedia Projects). It contained two trillion tokens with the current division in five collections<sup>21</sup>. The current version of Common Corpus, the one we describe in this paper, was released in February 2025 for the Paris AI Summit. Common Corpus will remain an evolving resource reflecting the need for a continuous open infrastructure rather than a fixed dataset, with HuggingFace ensuring a proper archiving of past versions for reproducibility.

A primary driver of Common Corpus has been what we term the **open data paradox**: most of the content available under free licenses or uncopyrighted is not integrated into the existing pretraining datasets. Since GPT-3, the pretraining data infrastructure has been extensively reliant on web archiving and the constraints that stem from it. which include the lack of available tools for PDF hosting and processing, or for diverse language detection.

The open data paradox had the effect of making the open pretraining ecosystem viable in the first place. In particular, the gradual filtering of web archives according to more and more ethical rules leaves only a tiny residual amount under open licenses (less than 2% of the original set).

## B Provenance

### B.1 Open Government

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.

#### B.1.1 Finance Commons

The datasets that make up Finance Commons<sup>22</sup> are presented in Table 4. 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<sup>23</sup>, compiled using the EDGAR-Crawler toolkit<sup>24</sup>.

<sup>19</sup>Released as <https://huggingface.co/datasets/PleIAs/French-PD-Books> and <https://huggingface.co/datasets/PleIAs/French-PD-Newspapers>.

<sup>20</sup>Release blogpost: <https://huggingface.co/blog/Pclanglais/common-corpus>

<sup>21</sup>Release blogpost: <https://huggingface.co/blog/Pclanglais/two-trillion-tokens-open>

<sup>22</sup><https://huggingface.co/collections/PleIAs/finance-commons-66925e1095c7fa6e6828e26c>

<sup>23</sup><https://www.sec.gov/search-filings/edgar-search-assistance/accessing-edgar-data>

<sup>24</sup><https://github.com/nlpaeub/edgar-crawler>

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

Table 4: Finance Commons sources distribution with languages.

- **World Trade Organization (WTO).** This dataset comprises documents from WTO’s official Documents Online platform<sup>25</sup>. 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<sup>26</sup>.
- **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<sup>27</sup> (AMF), which is an independent public authority that regulates the French market. The documents are primarily in French. Also released separately as AMF-PDF<sup>28</sup>.
- **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<sup>29</sup>, 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 4).
- **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 available by the General Council of the WTO in 2006<sup>30</sup>. 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.

### B.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<sup>31</sup>. It includes texts from 21 EU languages. It was originally compiled by Koehn (2005).

<sup>25</sup>[https://docs.wto.org/dol2fe/Pages/FE\\_Search/FE\\_S\\_S005.aspx](https://docs.wto.org/dol2fe/Pages/FE_Search/FE_S_S005.aspx)

<sup>26</sup><https://huggingface.co/datasets/PleIAs/WTO-PDF>

<sup>27</sup><https://www.amf-france.org/en/news-publications/publications/open-data>

<sup>28</sup><https://huggingface.co/datasets/PleIAs/AMF-PDF>

<sup>29</sup><https://ted.europa.eu/en/>

<sup>30</sup>[https://www.wto.org/english/docs\\_e/gattdocs\\_e.htm](https://www.wto.org/english/docs_e/gattdocs_e.htm)

<sup>31</sup><https://www.statmt.org/europarl/>

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

Table 5: Legal Commons sources distribution with languages.

- **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<sup>32</sup>. The dataset spans the years 1658 to 2020.
- **CourtListener.** This is a dataset<sup>33</sup> of opinions, oral arguments, judges, judicial financial records, and federal filings put together by the Free Law Project<sup>34</sup>.
- **EUR-lex.** This is a dataset of 57,000 legislative documents from the EU<sup>35</sup>. It is based on the dataset by Loza Mencía and Fürnkranz (2010) and developed by Chalkidis et al. (2019). The documents have also been annotated by the Publications Office of EU<sup>36</sup> with concepts from EuroVoc<sup>37</sup>. 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<sup>38</sup>, which is a data repository for the Publications Office of the European Union. This dataset was originally compiled by Sébastien Campion<sup>39</sup>.
- **French Open Data.** This dataset comes from French administrative bodies’ websites, for example, the French Directorate of Legal and Administrative Information

<sup>32</sup><https://case.law/>

<sup>33</sup><https://www.courtlistener.com/help/api/bulk-data/>

<sup>34</sup><https://free.law/contact>

<sup>35</sup><https://eur-lex.europa.eu/>

<sup>36</sup><https://publications.europa.eu/en>

<sup>37</sup><http://eurovoc.europa.eu/>

<sup>38</sup><https://op.europa.eu/en/web/cellar>

<sup>39</sup><https://huggingface.co/datasets/EuropeanParliament/Eurovoc>

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

Table 6: Subsets of Open Culture with language coverage, domains, and token counts.

(Direction de l’information légale et administrative<sup>40</sup>; 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)<sup>41</sup>.
- **UN Digital Library.** This dataset comes from the UN Digital Library<sup>42</sup>.
- **PleIAs European Legal Dataset.** We also collect datasets from various EU websites, *e.g.*, Archives of the EU Institute<sup>43</sup> and the Council of the EU<sup>44</sup>.
- **OECD.** These data come from the Organisation for Economic Co-operation and Development (OECD)<sup>45</sup>.

## B.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:

<sup>40</sup><https://echanges.dila.gouv.fr/OPENDATA/>

<sup>41</sup><https://huggingface.co/datasets/pile-of-law/pile-of-law>

<sup>42</sup><https://digitallibrary.un.org/?ln=en>

<sup>43</sup><https://archives.eui.eu/>

<sup>44</sup><https://www.consilium.europa.eu/en/general-secretariat/corporate-policies/transparency/open-data/>

<sup>45</sup><https://www.oecd.org/en/data/datasets.html?orderBy=mostRelevant&page=0>

- **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<sup>46</sup>.
- **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 6). 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<sup>47</sup>. 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).
- **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<sup>48</sup>. 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<sup>49</sup>. The newspapers were digitized by the Library of Congress. The dataset can be fully explored through an original corpus map created by Nomic AI<sup>50</sup>. 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<sup>51</sup>. 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 published between 1618 and 1990 (Neudecker, 2016). The original sources are available via Europeana, and were made available by Big Science<sup>52</sup>.
- **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<sup>53</sup>. The articles were originally published between 1794 and 1957. This dataset was curated and first made available by Sebastian Majstorovic<sup>54</sup>.

---

<sup>46</sup><https://api.bnf.fr>

<sup>47</sup><https://shiny.ens-paris-saclay.fr/app/gallicagram>

<sup>48</sup><https://www.loc.gov/collections/selected-digitized-books/about-this-collection/>

<sup>49</sup><https://chroniclingamerica.loc.gov/>

<sup>50</sup><https://atlas.nomic.ai/data/aaron/pdnews-21286k-tr2k-addmeta/map>

<sup>51</sup><https://paperspast.natlib.govt.nz/newspapers>

<sup>52</sup>[https://huggingface.co/datasets/biglam/europeana\\_newspapers](https://huggingface.co/datasets/biglam/europeana_newspapers)

<sup>53</sup><https://www.deutsche-digitale-bibliothek.de/newspaper>

<sup>54</sup><https://huggingface.co/datasets/storytracer/German-PD-Newspapers>

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 7: Token count by dataset Open Science.

- **German PD.** This dataset contains texts from various sources, including the Mannheim Corpus of Historical Newspapers and Magazines<sup>55</sup> (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<sup>56</sup> (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<sup>57</sup>. 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<sup>58</sup>. OCR was done using Nautilus-OCR<sup>59</sup>. 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.

### B.3 Open Science

In Table 7, we present the total token counts per collection inside of the Open Science part of Common Corpus.

### B.4 Open Code

Table 8 shows the number of tokens for the top ten coding languages and frameworks in Open Code.

<sup>55</sup><https://repos.ids-mannheim.de/fedora/objects/clarin-ids:mkhz1.00000/datastreams/CMDI/content>

<sup>56</sup><https://www.bne.es/fr/catalogues/biblioteca-digital-hispanica>

<sup>57</sup><https://www.digitisednewspapers.net/histories/delpher/>

<sup>58</sup><https://data.bnlu/>

<sup>59</sup><https://github.com/natliblalu/nautilusocr>

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

Table 8: Token counts by programming language or framework.

## C Open Culture Verification

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

- **Author life + 70 years for all non-US authors.** Among most signatories of the Berne Convention for the Protection of Literary and Artistic Works<sup>60</sup>, 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.

## D Cleaning and Curation

### D.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 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

---

<sup>60</sup><https://www.wipo.int/treaties/en/ip/berne/>

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).

## D.2 OCR Error Detection

**OCRosope.** 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 le Uc-régeik a des craintes ; il a fait venir chez lui les membres du tribunal de la cour , et leur a remis son testament. La fermentation qu'a causée l'arrestation de Thorild n'est pas appaisée par le langage qu'il a utilisé au duc-régent a été bien entendu par le peu) le y ir M» (Qui n'entendrait pas l'apostrophe suivante ? ttRxc3xa7nd >la libuk à autre raison , et ne nous force pas de l'achever i te n :e sang.,

Le duc a fait exécuter sur-le-champ une femme à la prison de la ville , pour les détourner de leur laisser se faire par de faux bruits et des jugemens prévus , et en même temps l'ordre a été donné à la garnison de charger et de se tenir prêts 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 her, but turned away and passed from her presence, filled with sorrow and moved with sympathy, which her intense emotions seemed to communicate to even these thoughtless young men of the plains.

And the OCRerrcr detection (with formatting for clarity):

They did not approach <er>her,</er> but turned away and passed <er>from</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>the</er> plains.

### D.3 OCR Correction

Here is an example of text containing various OCR errors:

The guaranties 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.

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

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.