The German Commons – 154 Billion Tokens of Openly Licensed Text for German Language Models

Lukas Gienapp
University of Kassel, hessian.AI, and ScaDS.AI
Kassel, Germany

Christopher Schröder
InfAI and ScaDS.AI
Leipzig, Germany

Stefan Schweter
Independent Researcher
Holzkirchen, Germany

Christopher Akiki
Leipzig University and ScaDS.AI
Leipzig, Germany

Ferdinand Schlatt
Friedrich-Schiller-Universität Jena
Jena, Germany

Arden Zimmermann
German National Library
Leipzig, Germany

Philippe Genêt
German National Library
Frankfurt, Germany

Martin Potthast
University of Kassel, hessian.AI, and ScaDS.AI
Kassel, Germany

Abstract

Large language model development relies on large-scale training corpora, yet most contain data of unclear licensing status, limiting the development of truly open models. This problem is exacerbated for non-English languages, where openly licensed text remains critically scarce. We introduce the *German Commons*, the largest collection of openly licensed German text to date. It compiles data from 41 sources across seven domains, encompassing legal, scientific, cultural, political, news, economic, and web text. Through systematic sourcing from established data providers with verifiable licensing, it yields 154.56 billion tokens of high-quality text for language model training. Our processing pipeline implements comprehensive quality filtering, deduplication, and text formatting fixes, ensuring consistent quality across heterogeneous text sources. All domain subsets feature licenses of at least CC-BY-SA 4.0 or equivalent, ensuring legal compliance for model training and redistribution. The *German Commons* therefore addresses the critical gap in openly licensed German pretraining data, and enables the development of truly open German language models. We also release code for corpus construction and data filtering tailored to German language text, rendering the *German Commons* fully reproducible and extensible.

Data: https://huggingface.co/datasets/coral-nlp/german-commons
Code: https://github.com/coral-nlp/llmdata

1 Introduction

Open language models are increasingly rivaling commercial systems in terms of their effectiveness and/or efficiency on key benchmarks, with expanding coverage of languages and tasks. Yet, the degree of openness of open models is often lacking [1]. While the weights of open models are being released under open licenses, the licensing of the training data of many models remains unclear. However, as most pretraining datasets used are derived from large-scale web crawls, this creates legal and ethical barriers to the development of fully open language models, despite the web's long-established value in science [2] and industry [3]. This is because (1) the provenance of web content is hard to establish and (2) obtaining consent from original authors or copyright holders is infeasible at web scale; (3) re-publishing web data without such consent as part of a training dataset infringes upon copyright and privacy; in particular since (4) the data may contain personally identifiable information (PII), and (5) generative models, even if published openly, may reproduce sensitive or copyrighted text.

Table 1: Overview of number of documents, number of tokens and median sequence length per source domain comprised in the German Commons corpus.

German Commons				
Domain	**Documents**		**Tokens**	
🌐 Web	15.48 M	43.26 %	19.89 B	12.87 %
💬 Political	0.26 M	0.72 %	3.57 B	2.31 %
⚖️ Legal	0.51 M	1.44 %	2.99 B	1.94 %
📰 News	13.27 M	37.08 %	72.67 B	47.02 %
🏦 Economics	0.06 M	0.16 %	0.11 B	0.07 %
📚 Cultural	6.11 M	17.08 %	54.49 B	35.25 %
🔬 Scientific	0.09 M	0.26 %	0.84 B	0.54 %
∑ **Total**	35.78 M		154.56 B	

These shortcomings limit the usefulness of many open-weight models. Practitioners can only trust, but not verify, that, for example, the web crawls used for training have not been contaminated by benchmark test data or that no copyrighted material is included. To minimize these risks and make open language models usable for both scientific and commercial purposes without reservations, it is important to limit their training data to verifiably open texts. However, this is challenging for non-English languages, as the available open alternatives to web data need to be carefully compiled.

We therefore introduce the *German Commons*, the largest pretraining-scale collection of explicitly openly licensed text in German language (Table 1). It encompasses 154.56 billion tokens of open text across 35.78 million documents spanning seven thematic domains. This renders it the largest openly licensed German corpus (Section 2 and 3). Alongside the corpus, we release our data processing library *llmdata* to ensure full reproducibility (Section 4). Moreover, we provide an analysis of the corpus' properties (Section 5). In the Appendix, a datasheet compliant with the recommendation of Gebru et al. [4] is included.

2 Related Work

To motivate our choices for constructing the *German Commons*, we revisit three categories of existing training corpora: (1) the web-scraped datasets that dominate current LLM training, (2) smaller-scale, German-specific resources, and (3) emerging openly licensed alternatives, predominantly in English.

Web Corpora for LLM Training Modern LLM training relies on web-scraped content as large-scale text source, with collections such as *C4* [5]/*mC4* [6], *The Pile* [7], *OSCAR* [8], *ROOTS* [9], *RedPajama* [10], *Dolma* [11], *HPLT* [12], and *FineWeb* [13, 14]. However, these datasets derive almost exclusively from Common Crawl[1], which creates dependency on a single source and lacks explicit licensing metadata. While C4Corpus [15] identifies licensed content through substring matching, license scope remains unclear—content may be *open* but not *verifiable*. Additional risks include terms of service restrictions that prohibit model training [16] and a considerable amount of PII despite preprocessing [17]. The heterogeneous nature of web data further necessitates extensive quality filtering [11, 18, 19]. This creates a fundamental tension: web-scraped corpora provide scale but introduce legal, ethical, and quality risks. The *German Commons* aim to address these shortcomings by collecting verifiably licensed, high-quality text content from non-web sources.

German Text Datasets Web datasets naturally include German subsets, however, with identical licensing and quality issues. Several additional large-scale German text corpora are available, but mostly predate LLM training efforts, such as the *Leipzig Corpora Collection* [20], *OPUS* [21], and *HPLT* [12]. Moreover, since they are also sourced from the web, their licenses are frequently unclear and not verifiable as well. While openly licensed German corpora exist (see Section 3), their individual volumes are substantially smaller than web-scraped alternatives, and they remain fragmented without centralized access. Furthermore, multiple of these datasets are not available in plain text form, and thus need to undergo text extraction and preprocessing first. The *German Commons* aim to instead provide a unified, comprehensive source of text data suitable for large language model training.

Openly Licensed Training Corpora Scaling openly licensed text corpora faces verification challenges, leading datasets to include questionable content [1]. While code datasets can rely on the explicit machine-readable licensing present in code repositories [22], such annotations are rarely available for natural-language text from web or print sources, which consequently proves more difficult to verify. Therefore, data collection efforts turn to trusted providers of licensed and open-domain content, such as government agencies, GLAM institutions, and collaborative projects such as Wikimedia. Current large-scale openly-licensed collections for language model training include: (1) the *Open License Corpus* [23] is an aggregation from existing corpora covering open domain and attribution licensed content from the legal, scientific, conversational, books, and news domain, amounting to 228B tokens of multi-domain English text; (2) the *KL3M* project [16] assembled 1.35 trillion tokens of English text sourced primarily from open domain government records in English language; it explicitly excludes content licenses with a 'share-alike' clause such as CC-BY-SA and thus removes, e.g., Wikipedia from consideration. (3) the *Common Pile* [24] compiles approximately upwards of two trillion tokens of English text with stringent licensing requirements, and is the largest, yet monolingual resource of verifiably open text; and (4) the *Common Corpus* [25] also provides 2 trillion multilingual tokens, including 112 billion German tokens, which serve as a the starting point for our data collection efforts.

[1] https://commoncrawl.org/

3 Sourcing Open German Text Data

This section provides our working definition of 'openly licensed' and detailed provenance of all data included in the *German Commons*. Dataset construction begins with the German subset of the existing *Common Corpus* [25], applying stricter licensing and quality criteria that reduce usable tokens from 112B to 70B. We then expand coverage through updated source dataset versions and previously unconsidered collections, combining existing resources with newly assembled open data. Table 2 details all constituent datasets with source, license, type, and size statistics. Data collection uses the newest available versions through August 31st, 2025.

3.1 Where do we obtain data from?

We identify two source types for *German Commons*: (1) established open corpora with published full texts; and (2) (metadata) collections requiring text extraction from source files. We use provided texts where available, otherwise crawling and extracting plain text from sources. Data integration spans multiple providers: Text+, Zenodo, Huggingface, German National Library (DNB), Austrian National Library (ÖNB), German Digital Dictionary (DWDS), Leibniz-Institute for German Language (IDS), and Wikimedia projects.[2] From their catalogs, we obtain all primarily German collections with explicit open licenses. For the non-curated platforms Zenodo and Huggingface, we select only uploads from trusted parties with clear licensing protocols.

3.2 What do we consider openly licensed?

The *German Commons* require explicit licenses for each document. We exclude datasets with ambiguous licensing, i.e., where the aggregation or metadata carries open licenses while underlying text content retains unspecified copyright restrictions. This excludes most web-crawled datasets that conflate aggregation with content licensing rights [24].

We follow Kandpal et al. [24] in adopting the Open Knowledge Foundation's Open Definition 2.1[3]. Unlike *Open License Corpus* and *KL3M*, this covers licenses with a share-alike provision. Table 3 lists the accepted licenses found in our data. Licenses are grouped into the categories of (1) public domain equivalent (2) attribution

[2]

Text+	https://text-plus.org
Zenodo	https://zenodo.org
Huggingface	https://huggingface.co/datasets
German National Library (DNB)	https://dnb.de
Austrian National Library (ÖNB)	https://www.onb.ac.at
German Digital Dictionary (DWDS)	https://www.dwds.de/
Leibniz-Institute for German Language (IDS)	https://www.ids-mannheim.de
OpenAlex	https://openalex.org
Wikimedia	https://www.wikimedia.de

[3] https://opendefinition.org/od/2.1/en/

Table 2: Overview on datasets constituting the German Commons corpus. Token counts are measured using GPT-2 tokenizer. 'Various' licenses are open, but differ per document as per original source.

Thematic Subset & Constituent Corpora	Ref.	License	Text Type	Docs (# / %)		Tokens (# / %)	
Web Commons				15.48 M	43.26 %	19.89 B	12.87 %
Wikipedia	[26]	CC-BY-SA-4.0	Various	2.93 M	8.19 %	2.95 B	1.91 %
Wikivoyage	[27]	CC-BY-SA-4.0	Travel	0.02 M	0.06 %	0.04 B	0.03 %
Wikipedia Discussions	[28]	CC-BY-SA-4.0	Online Discussions	8.35 M	23.34 %	1.22 B	0.79 %
Youtube-Commons	[25]	Various	Video Subtitles	2.81 M	7.85 %	14.48 B	9.37 %
One Million Posts Corpus	[29]	CC-BY-4.0	Online Discussions	0.95 M	2.64 %	0.09 B	0.06 %
The Stack (Markdown and TXT Subsets)	[22]	Various	Various	0.42 M	1.18 %	1.11 B	0.72 %
Political Commons				0.26 M	0.72 %	3.57 B	2.31 %
Reichtagsprotokolle	[30]	CC-BY-SA-4.0	Parliamentary Protocols	0.00 M	0.00 %	0.70 B	0.46 %
German Political Speeches	[31]	CC-BY-4.0	Speech Transcripts	0.01 M	0.02 %	0.03 B	0.02 %
Corpus der Drucksachen des Deutschen Bundestages	[32]	CC0-1.0	Parliamentary Publications	0.00 M	0.01 %	0.53 B	0.34 %
C. d. Plenarprotokolle des Deutschen Bundestages	[33]	CC0-1.0	Parliamentary Protocols	0.00 M	0.01 %	0.32 B	0.20 %
EuroVoc		EUPL	Parliamentary Publications	0.25 M	0.69 %	1.99 B	1.29 %
Legal Commons				0.51 M	1.44 %	2.99 B	1.94 %
Corpus des Deutschen Bundesrechts	[34]	CC0-1.0	German Federal Laws	0.00 M	0.01 %	0.00 B	0.00 %
OpenLegalData	[35]	CC0-1.0	Court Decisions	0.23 M	0.70 %	1.92 B	1.24 %
Corpus der Entscheidungen des BFH	[36]	CC0-1.0	Court Decisions	0.01 M	0.03 %	0.07 B	0.04 %
Entscheidungen des BGH in Strafsachen des 20. Jhd.	[37]	CC0-1.0	Court Decisions	0.04 M	0.10 %	0.09 B	0.06 %
Corpus der Entscheidungen des BGH	[38]	CC0-1.0	Court Decisions	0.08 M	0.22 %	0.29 B	0.19 %
Corpus der Entscheidungen des BVerfG	[39]	CC0-1.0	Court Decisions	0.01 M	0.02 %	0.04 B	0.03 %
Corpus der Entscheidungen des BpatG	[40]	CC0-1.0	Court Decisions	0.03 M	0.09 %	0.19 B	0.12 %
Corpus der Entscheidungen des BVerwG	[41]	CC0-1.0	Court Decisions	0.03 M	0.08 %	0.12 B	0.08 %
Corpus der amtl. Entscheidungssammlung des BVerfG	[42]	CC0-1.0	Court Decisions	0.00 M	0.00 %	0.02 B	0.02 %
Corpus der Entscheidungen des BAG	[43]	CC0-1.0	Court Decisions	0.01 M	0.02 %	0.05 B	0.03 %
EurLEX	[44]	CC-BY-4.0	European Union Laws	0.06 M	0.18 %	0.20 B	0.13 %
News Commons				13.27 M	37.08 %	72.67 B	47.02 %
Deutsches Zeitungsportal		CC0-1.0	News Articles	8.08 M	22.57 %	43.87 B	28.38 %
Europeana Newspapers		CC0-1.0	News Articles	3.26 M	9.10 %	20.68 B	13.38 %
ANNO		CC0-1.0	News Articles	1.91 M	5.34 %	8.10 B	5.24 %
Wikinews		CC-BY-SA-4.0	News Articles	0.02 M	0.07 %	0.01 B	0.01 %
Economics Commons				0.06 M	0.16 %	0.11 B	0.07 %
TEDEUTenders	[25]	CC0-1.0	Procurement Notices	0.06 M	0.16 %	0.11 B	0.07 %
Cultural Commons				6.11 M	17.08 %	54.49 B	35.25 %
DiBiLit-Korpus	[45]	CC-BY-SA-4.0	Literature	0.00 M	0.01 %	0.22 B	0.14 %
DiBiPhil-Korpus		CC-BY-SA-4.0	Literature	0.00 M	0.00 %	0.03 B	0.02 %
Wikisource		CC-BY-SA-4.0	Various	0.24 M	0.67 %	0.35 B	0.23 %
German-PD	[25]	CC0-1.0	Literature	0.12 M	0.35 %	49.33 B	31.92 %
BLBooks	[46]	CC0-1.0	Literature	0.00 M	0.01 %	1.01 B	0.65 %
MOSEL	[47]	CC-BY-4.0	Speech Transcripts	3.13 M	8.74 %	3.18 B	2.06 %
SBB Fulltexts	[48]	CC-BY-4.0	Literature	2.61 M	7.28 %	0.36 B	0.23 %
Wikiquote		CC-BY-SA-4.0	Quotes & Proverbs	0.01 M	0.02 %	0.01 B	0.00 %
Scientific Commons				0.09 M	0.26 %	0.84 B	0.54 %
Wikibooks	[49]	CC-BY-SA-4.0	Educational Books	0.03 M	0.08 %	0.05 B	0.03 %
Digitalisierung des Polytechnischen Journals	[50]	CC-BY-SA-4.0	Scholarly Papers	0.00 M	0.00 %	0.18 B	0.12 %
Directory of Open Access Books		Various	Scholarly Books	0.00 M	0.01 %	0.17 B	0.11 %
arXiv		Various	Scholarly Papers	0.00 M	0.00 %	0.00 B	0.00 %
Wikiversity		CC-BY-SA-4.0	Educational Content	0.02 M	0.05 %	0.03 B	0.02 %
OpenALEX	[51]	Various	Scholarly Papers	0.05 M	0.13 %	0.41 B	0.27 %
Total				35.78 M		154.56 B	

Table 3: Licenses of data constituting *German Commons*.

	License		Attribution ↓	Copyright Notice ↓	Share-Alike ↓	Source Provision ↓
PD-Equiv.	CC0-1.0 / Public Domain	🔗				
	Unlicense	🔗				
	MIT-0	🔗				
	0BSD	🔗				
Attribution	MIT	🔗	✓	✓		
	BSD-2-Clause	🔗	✓	✓		
	BSD-2-Clause-FreeBSD	🔗	✓	✓		
	BSD-3-Clause	🔗	✓	✓		
	BSD-Source-Code	🔗	✓	✓		
	Apache-1.1	🔗	✓	✓		
	Apache-2.0	🔗	✓	✓		
	BSD-4-Clause-UC	🔗	✓	✓		
	BSD-4-Clause	🔗	✓	✓		
	CC-BY-2.0	🔗	✓	✓		
	CC-BY-3.0	🔗	✓	✓		
	CC-BY-4.0	🔗	✓	✓		
Copyleft	CC-BY-SA-4.0	🔗	✓	✓	✓	
	EUPL-1.2	🔗	✓	✓	✓	✓
	Artistic-2.0	🔗	✓	✓	✓	✓

licenses; and (3) copyleft licenses. All of the selected licenses permit redistribution, modification, and commercial use. They thus support data commons principles for sustainable open model development [52, 53]. The latter two require attribution and/or license indication, while copyleft licenses have to be redistributed under the same license terms (share-alike). Each document in the *German Commons* is tagged with its corresponding SPDX-canonical license[4], linking to its original license text. We exclude licenses with non-commercial clauses, research-only provisions, and other use-limiting conditions. However, we advice practitioners to review individual license compatibility before use.

We acknowledge that this approach, i.e., trusting provided licenses without independent auditing, carries inherent misattribution risks. However, given the relative scarcity of open German text, and the institutional nature of the data providers contributing to the *German Commons*, we consider it a viable method for large-scale data collection. While we apply less strict criteria than Kandpal et al. [24], who independently audit entries in their data, to reasonably mitigate risks, we consequently limit inclusion of data to established institutional providers: national libraries, academic institutions, government agencies, and verified open-source platforms, excluding sources lacking clear licensing protocols.

[4] https://spdx.org/licenses/

3.3 Detailed Provenance Information

The *German Commons* are divided into seven thematic domains: *Web* spans collaborative platforms, user-generated content and discussions, and open-source repositories; *Political* encompasses parliamentary protocols, publications, and speeches; *Legal* includes court decisions, proceedings, and regulatory frameworks from German and European judicial systems. While previous datasets group political and legal text as one [24], we argue that legal text has a distinct writing style not overlapping with the more general text published by political institutions and thus chose to differentiate between both. The *News* content draws primarily from historical newspaper archives maintained by cultural heritage institutions. *Economics* includes documents from business and trade publications, *Cultural* includes books and general speech transcripts, and *Scientific* includes scholarly and educational content. The constituting data of each domain is detailed in the following paragraphs.

Web Commons For *Wikipedia* and *Wikivoyage*, we use the TEI-encoded version supplied by the DWDS [26, 27]. The complementary corpus *Wikipedia Discussions* of user discussions on Wikipedia is supplied by the IDS [28]. The *One Million Posts Corpus* consists of user comments posted to the Austrian newspaper 'Der Standard', collected and openly licensed in collaboration with the newspaper [29]. *Youtube-Commons* [25] encompasses audio transcripts of over 2 million videos shared on YouTube under a CC-BY license. These include both automatically translated and manually curated texts. Finally, we filter German-language website sources and README files hosted on GitHub by filtering the Markdown and TXT subsets of *The Stack* [22] corpus for our allowed licenses.

Political Commons We aggregate official publications from the German federal parliament (*Corpus der Drucksachen des Deutschen Bundestags*, [32]) and the EU parliament (*EuroVoc*). Alongside that, we include parliamentary protocols and speeches from the German federal parliament (*Corpus der Plenarprotokolle des Deutschen Bundestags*, [33]) and the historic Reichstag (*Reichtagsprotokolle*, [30]). German parliamentary documents are exempt from copyright as official works. EU documents are licensed under the European Union Public License (EUPL), which is compatible with a CC BY-SA 3.0 license.[5] Additionally, we include a corpus of political speeches in German language [31].

Legal Commons The largest portion of legal documents comprises court proceedings from German courts, collected by *OpenLegalData* [35].[6] Proceedings of federal courts are made available in dedicated corpora, namely for the Bundesfinanzhof (*BFH*, [36]), Bundesgerichtshof (*BGH*, [37, 38]), Bundesverfassungsgericht (*BVerfG*, [39, 42]), Bundespatentgericht (*BPatG*, [40]), Bundesverwaltungsgericht (*BVerwG*, [41]), and Bundesarbeitsgericht (*BAG*, [43]). Court decisions and decrees of German courts are exempt from copyright. Additionally, we include European laws made available via the *EUR-Lex*[7] platform, incorporating the official data dump.

[5] https://eupl.eu/1.2/en
[6] https://openlegaldata.io
[7] https://eur-lex.europa.eu/

News Commons The *Deutsches Zeitungsportal*[8] corpus covers over 4 million German newspaper editions of nearly 2000 newspapers from 1671 to 1994, incorporating all public domain releases. The *ANNO* corpus provides equivalent coverage for Austrian newspapers, spanning around 1600 newspapers in the public domain. Additionally, we include the *Europeana Newspapers* archive, which contains more than 4 million individual documents across multiple European languages from the 18th to early 20th century, using the plain text version created by the BigLAM initiative.[9] *Wikinews* is a wiki project for news articles written by community members; we obtain a current dump and parse plain text from it.

Economics Commons The *TEDEUTenders* [25] dataset comprises European public procurement notices from the online version of the EU's Official Journal Supplement, providing structured access to public tender information across EU member states.

Cultural Commons The *DiBiLit* [45] dataset provides a literary corpus spanning the 16th through early 20th centuries, containing primarily German literary works (poetry, drama, prose) alongside select humanities texts. *DiBiPhil* covers the 15th to 20th centuries, encompassing literary works with philosophical content. Both corpora feature TEI-encoded, homogenized texts. *Wikisource* is a collaborative digital library for public domain source texts, including historical documents, literary works, and reference materials, transcribed and proofread by volunteer contributors. *GermanPD* [25] systematically aggregates German monographs and periodicals from Internet Archive and European national libraries, including both OCR-sourced content and text extracted from machine-readable format. *BLBooks* [46] comprises digitized pages from the British Library, primarily concentrated on the 18th-19th century and covering diverse subject areas. While BLBooks is originally split into pages, we concatenate pages of the same in the provided order to re-assemble complete texts. *MOSEL* [47] aggregates multilingual open-source speech recordings with Whisper-generated transcriptions across 24 EU languages. *SBB Fulltexts* [48] is a collection of the fulltexts available in the digitized collections of the Berlin State Library (SBB), covering approximately 25 000 unique German literary works split into individual page scans. *Wikiquote* collaboratively collects quotations and proverbs that fall under public domain.

Scientific Commons The *Digitalisierung des Polytechnischen Journals* is a historic technical periodical, digitized through OCR [50]. *Wikibooks* is a free repository of open-content textbooks and educational materials, converted to TEI [49]. We crawl the *Directory of Open Access* books (DOAB) for scholarly open-access book publications and filter for German language, similar to the English counterpart assembled by Kandpal et al. [24]. We further include German scholarly articles published on *arXiv* which explicitly indicate open licensing. For each article, we only include the latest version to reduce deduplication overhead. Finally, we obtain all fulltexts from *OpenAlex* [51], a metadata aggregator for open-access scholarly papers, and filter for German texts under open licenses. Since OpenAlex does not provide fulltexts, we instead crawl all linked PDFs and extract plain text from those.

[8] https://www.deutsche-digitale-bibliothek.de/newspaper
[9] https://huggingface.co/biglam

4 Data Processing

We apply several processing steps to transform the heterogeneous input data into a consistent output format, apply quality heuristics and filters, and fix text formatting. The individual steps are detailed in this section and are executed in the listed order. We include detailed statistics on removed documents and tokens per processing step and source dataset in Appendix A.4. Table 4 shows the final dataset schema. The dataset is distributed in Parquet format, with partial files partitioned for each thematic domain and source dataset to allow selective loading.

Plain Text Extraction Most of the compiled source data sets already feature readily available plain text. In instances where the original corpus is only available in PDF format, we use `Grobid` [54] (for scholarly sources) or `OlmOCR` [55] (for other types of PDF) to obtain plain text. In sources where TEI or similar formats featuring structural markup are used, we convert to plain text and systematically exclude editorial elements such as title pages, page numbering and breaks, footnotes, bibliographic references, and textual apparatus, preserving only the core textual content. Markdown syntax is left intact when encountered. For wiki markup, we use `mwparserfromhell`[10] to convert to plain text.

Text Formatting To address artifacts introduced during optical character recognition present in a large portion of the data, we further apply a minimal amount of text formatting. We apply the standard formatters of the FTFY suite [56], including UTF-8 encoding fixes, removal of HTML entities, control characters and terminal escape sequences, ligature decomposition, character width and surrogate pair fixes, quote character normalization, and Unicode NFC normalization. In addition, we apply a series of custom regex-based transformations to address common whitespace issues resulting from OCR-sourced data. It collapses multiple spaces and tabs into single spaces, reduces excessive newlines while preserving paragraph boundaries, removes hyphenation on line breaks, and removes leading and trailing whitespace from individual lines.

Language & Length Filtering For datasets that indicate language in their metadata, we pre-filter using each dataset's own language tags to reduce redundant classification work. Then, we employ the FastText language identification model [57] to automatically detect the language of the input texts. We apply the compressed model version [58], which supports 176 languages, to text snippets truncated to 4096 characters for computational efficiency. The classifier treats the input by replacing newlines with spaces to improve the prediction accuracy, as the FastText model was trained on single-line text samples. We discard text not classified as primarily German with a probability of at least 0.65.

We use the GPT-2 tokenizer [59] to obtain token counts for all sequences in the corpus. We discard sequences shorter than 32 tokens, as the majority of those are extraction artifacts. The token count is persisted alongside the text data for downstream filtering.

Quality Filtering Following the consideration of other large-scale pretraining datasets, such as the BigScience ROOTS corpus [9] for BLOOM [60], Gopher [61], and CulturaX [62], we implement several quality indicators to remove low-quality text. Those include

[10] https://github.com/earwig/mwparserfromhell

Table 4: Dataset schema

Key	Value
`id`	Document identifier as found in the original dataset
`source`	Source dataset this document stems from
`subset`	Thematic subset of this document
`text`	Cleaned full text of this document
`license`	Canonical SPDX license URL(s) of this document
`num_tokens`	Number of GPT-2 tokens in this document
`perplexity`	Perplexity of this document estimated by Wikipedia KenLM
`ocr_score`	OCR quality as calculated by `ocroscope`

word count, average word length, symbol-to-word ratios (hash and ellipsis), bullet/ellipsis line ratios, alphabetic word ratio, stop word count, text repetition through duplicate line/paragraph fractions, character-level duplicate fractions, and n-gram repetition analysis (2-4 grams for top frequency, 5-10 grams for duplication). We select the exclusion parameters for these indicators using percentile-based thresholds [62], calculating the value distributions on language-filtered and formatted data, and removing documents either below the 5th or above the 95th percentile, depending on indicator. As large subset of our data originates from OCR text, we additionally apply OCR-specific filtering heuristics to exclude errors not previously addressed through formatting fixes, namely character casing anomalies, fragmented words, and special character density. Exact parameters for all applied filters can be found in Appendix A.1. For all parameters, their choice was manually validated by cursory inspection of removed content for different percentile thresholds. Deviations from parameters used in prior work for English web-text [61] are little, and reasonable given the difference in language and domains of our data.

Deduplication For deduplication, we rely on the LSH bloom filter implementation of Dolma [11]. We perform paragraph-level deduplication, splitting each document into chunks at newline characters. MinHash collisions are detected using 20-gram shingling and a collision rate of 0.8, i.e., for two paragraphs to be deemed duplicates, 80% of their ngrams have to be identical. All but one chunk are removed from the corpus in case of collisions. The bloom filter was parametrized for a false positive rate of 1e-4. We make available the bloom filter file for others to deduplicate new data against *German Commons*.

PII Removal We remove personally identifiable information (PII) using a combination of regex-based filters and the Presidio framework [63]. Types of PII removed are email addresses, phone numbers, IP addresses, credit-card numbers, IBAN numbers, and URL. To keep the semantic structure of sentences intact, we replace each of them with respective generic information. A list of the replacements strings used is included in Appendix A.3, which allows lookup for full redaction or other replacements downstream.

License Mapping and Filtering We map the diverse license identifiers indicated in the data to a canonical set of SPDX license URLs pointing to the corresponding license of each document in the corpus. We then filter licenses to only open licenses as listed in Table 3. For cases where multiple licenses are given for a document, we require all of them to be permissive.

Table 5: Split Composition by domain for different training context lengths s. Assumes disjoint splits, i.e., $s \leq 8192$ contains all sequences $2048 < s \leq 8192$. Percentage for $\sum$ is in relation to full dataset.

	$s \leq 2048$		$s \leq 8192$		$s \leq 32768$		$s > 32768$	
📚	3.20 B	26.47 %	0.68 B	1.37 %	0.40 B	0.84 %	49.73 B	83.72 %
🏛	0.03 B	0.26 %	0.06 B	0.12 %	0.02 B	0.03 %	0.01 B	0.01 %
⚖	0.17 B	1.44 %	1.08 B	2.16 %	1.48 B	3.11 %	0.24 B	0.41 %
📰	3.27 B	27.05 %	39.17 B	78.60 %	30.24 B	63.38 %	0.00 B	0.00 %
💬	0.14 B	1.17 %	0.27 B	0.55 %	0.51 B	1.06 %	2.64 B	4.44 %
🔬	0.04 B	0.30 %	0.13 B	0.25 %	0.19 B	0.39 %	0.49 B	0.82 %
🌐	5.23 B	43.31 %	8.45 B	16.95 %	14.88 B	31.18 %	6.30 B	10.61 %
$\sum$	12.07 B	7.14 %	49.83 B	29.48 %	47.71 B	28.23 %	59.40 B	35.15 %

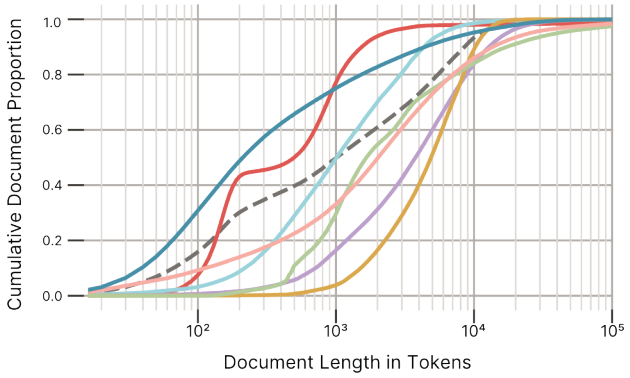

Figure 1: Cumulative proportion of tokens by document length in corpus, normalized by domain; overall (dashed), and by subset for cultural, economic, legal, news, political, scientific, web.

5 Corpus Statistics

We analyze corpus composition across thematic subsets and license types, demonstrate the efficacy of our filtering pipeline, and investigate different text properties, in order to illustrate *German Commons*' suitability for language model pretraining.

Token and Document Distribution By Thematic Domain. Figure 1 shows cumulative document proportions by length across thematic domains in *German Commons*. Distinct length distributions are apparent: cultural and web content concentrate in shorter documents, with cultural content showing rapid cumulative growth below 1,000 tokens due to page-level book segmentation in some source corpora. News content exhibits substantially longer documents, while legal, scientific, and political domains occupy intermediate positions similar to the overall corpus trend. Table 5 quantifies the practical implications of these length distributions for language model training contexts. When partitioning documents into subsets of short ($s \leq 2048$), medium ($s \leq 8192$), long ($s \leq 32\,768$) and very long ($s > 32\,768$) context lengths, domain distribution varies across partitions. At short contexts, web content dominates with 43.31% of available tokens, followed by news (27.05%) and cultural content

Table 6: Number of documents and tokens per license type.

License Type	Documents		Tokens	
Public Domain	13.95 M	35.96 %	126.61 B	74.91 %
Attribution	12.97 M	33.44 %	34.48 B	20.40 %
Copyleft	11.87 M	30.60 %	7.93 B	4.69 %

Table 7: Number of tokens per license type and domain.

Subset	Public-Domain		Attribution		Copyleft	
Cultural	49.83 B	39.36 %	3.54 B	10.27 %	0.64 B	8.08 %
Economic	0.11 B	0.09 %	—	—	—	—
Legal	2.78 B	2.19 %	0.20 B	0.58 %	—	—
News	72.66 B	57.39 %	—	—	0.01 B	0.18 %
Political	0.84 B	0.66 %	0.03 B	0.09 %	2.68 B	33.84 %
Scientific	0.29 B	0.23 %	0.10 B	0.30 %	0.44 B	5.61 %
Web	0.10 B	0.08 %	30.60 B	88.76 %	4.15 B	52.29 %

(26.47%). For medium and long contexts, news articles provide the majority of tokens at 78.60% and 63.38% respectively, with web content ranging from to 16.95% to 31.18%. At very long contexts, most tokens (83.73%) originate from book-length documents of the cultural domain. Scientific, legal, and political domains maintain smaller representation across all context lengths, ranging from 0.5-3% each. **All thematic domains remain represented in each length partition, enabling sub- or oversampling at different context lengths to control model exposure toward text genres.**

Token and Document Distribution By License Type. Table 6 presents document counts and token volume across license categories. Public domain equivalent licenses dominate token count with 126.61B tokens (74.291% of corpus) despite representing only 35.96% of documents, reflecting substantially longer average sequence lengths. This length disparity stems primarily from news articles and cultural content under public domain licensing. Attribution-type licenses contribute 34.48B tokens (20.40% of corpus) across 33.4% of documents, while copyleft licenses provide 7.93B tokens (4.69% of corpus) from 30.60% of documents. With licenses being derived from source datasets, license types exhibit strong correlations with thematic domains. Public domain content concentrates heavily in cultural (39.36% of public domain tokens) and news domains (57.39%), with minimal representation in web content (0.08%). Attribution licenses are predominantly found in web content (88.76% of attribution tokens), followed by cultural content (10.27%). Copyleft licenses span web sources (52.29%, including share-alike licensed Wikimedia projects) and political text (33.84%, primarily from EUPL-licensed *Eurovoc*). **All thematic domains feature public domain data, and public domain data yields 75% of tokens.**

Filtering Statistics. Our data filtering pipeline removes noisy data in three sequential stages (also see Section A.4). (1) Quality filtering removed 46.41% of initial data, with the majority of that being non-German text eliminated from multilingual source corpora (e.g., *The Stack*, *arXiv*) and very short texts being removed (e.g., Wikipedia redirect pages and failed PDF extractions). (2) Deduplication only removes an additionaly 2.7% of text, concentrated in web and news corpora which exhibit both within-corpus and cross-corpus overlaps. Other domains showed minimal duplication. (3) Final license compliance and PII filtering removed negligible volumes (0.08%). Source corpora contained minimal personally identifiable information, occurring predominantly in web sources and economics content, while other domains required virtually no PII filtering. Overall retention reached 50.73% of input, which, however, is attributable to select multilingual corpora, with the majority of source corpora retaining between 70% and 95% of their content. **This aligns with our filtering objective of eliminating noise while maximizing text retention at consistent quality. Only trace amounts of duplicates and PII were present in source corpora and subsequently removed.**

Table 8: Proportion of toxicity degree and corresponding toxicity label across paragraph sample. Differences for domains are minimal. No values above 3 are observed.

Kind	Non-Toxic				Mildly Toxic
	0	1	2	3	≥ 4
Ability	0.99	0.00	0.00	0.00	0.00
Gender/Sex	0.98	0.01	0.01	0.00	0.00
Race/Origin	0.94	0.01	0.04	0.01	0.00
Religion	0.97	0.02	0.01	0.00	0.00
Violence	0.86	0.11	0.03	0.01	0.00

Text Properties. We employ four pretrained encoder-based German text classifiers[11] to assess text properties across on a stratified random sample of 10 000 paragraphs per source corpus (385 467 total, due to less paragraphs available in some sources).

The toxicity classifier grades content on 0-7 scales across five categories (ability, gender/sex, race/origin, religion, and violence), with scores 0-3 considered non-toxic. Results listed in Table 8 show no paragraphs scoring above 3 in any category, with on average 95% of paragraphs scoring 0. Remaining scores fall between 1-3, with negligible differences between thematic domains. **The *German Commons* is thus deemed to contain only minimal amounts of harmful or toxic content.**

Language complexity classification (Figure 2) identifies four grades: plain (2%), easy (3%), everyday (65%), and special language (30%). Figure 2 reveals expected domain variations: scientific content exhibits highest special language proportion (63.8%), while web content shows highest everyday language (81.4%), with similar distributions found in Political, Economic, and Legal domains. News content demonstrates intermediate complexity with balanced distribution across categories. Cultural is nearly evenly divided between everyday (52.2%) and special language (39.9%). **Overall, a balanced complexity distribution across domains enables learning across linguistic registers.**

[11] Toxicity [64]: https://hf.co/PleIAs/celadon
Sentiment [65]: https://hf.co/oliverguhr/german-sentiment-bert
Complexity: https://hf.co/krupper/text-complexity-classification
Fluency: https://hf.co/EIStakovskii/bert-base-german-cased_fluency

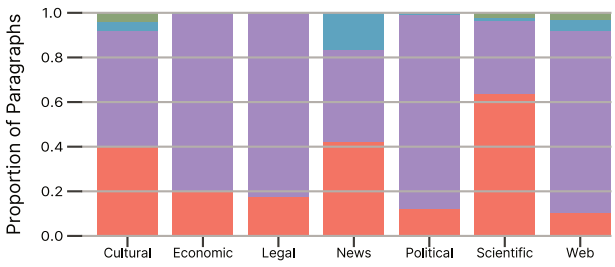

Figure 2: Proportion of text complexity (easy, simple, everyday, special) across paragraph sample, per subset.

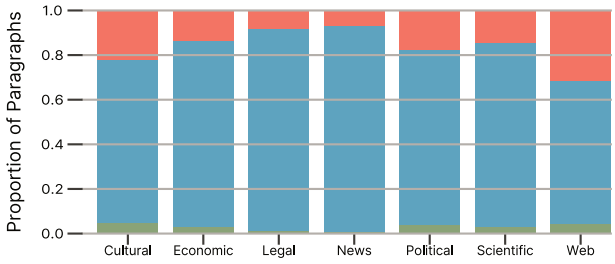

Figure 3: Proportion of text sentiment classes (negative, neutral, positive) across paragraph sample, per domain.

Sentiment analysis (Figure 3) categorizes text as negative (16.4%), neutral (80.5%), or positive (3.1%). Web content exhibits highest negative sentiment (31.8%), while cultural content shows most positive sentiment (5.0%). News content demonstrates the highest proportion of neutral sentiment (92.0%). The remaining domains maintain proportions similar to the overall distribution. **Mostly neutral text prevents systematic model biases w.r.t. sentiment.**

6 Limitations and Ethical Considerations

The German Commons inherits fundamental limitations from its constituent sources and curation methodology. We identify four primary limitations requiring explicit acknowledgment and propose future mitigation strategies.

First, the corpus exhibits temporal bias toward historical content. News (47.02%) and cultural domains (35.25%) comprise 82.27% of tokens, with cultural content predominantly sourced from 18th-20th century digitized texts. This historical skew induces nostalgia bias [66]. Scientific (0.54%) and economic (0.07%) domains remain critically underrepresented. Adding contemporary German text to rebalance the temporal distribution is paramount for future extensions of the corpus. Second, the predominance of OCR-extracted text may introduce errors. German diacritics exhibit heightened vulnerability to misrecognition [67]. While our pipeline applies OCR-specific filtering, residual errors may persist particularly in older texts. We did not apply LLM-based correction methods for error reduction, due to substantial computational expenditure, hallucination risks, and misinterpretation of historical texts, however, would be a future improvement if specialized correction models become available.

Third, standard German dominates content, diminishing linguistic diversity against non-standard varieties like Swiss, Austrian, or Low German dialects. Demographic biases may be present; socioeconomic stratification manifests through overrepresentation of formal registers from institutional sources. Cultural representation likely exhibits Western Protestant bias consistent with broader German NLP resources [68]. Targeted inclusion of dialectal and minority language varieties can improve this situation, if they become available under open licenses. Finally, privacy protection through PII removal provides limited security. Our regex- and Presidio-based approaches constitute surface-level modifications. However, the historical skew and data sources from the public record diminish the potential adverse effects should PII be contained in the data.

To enable informed downstream usage, we provide comprehensive documentation following established frameworks [4, 69] (Section A.4). The corpus includes document-level metadata preserving provenance and license information. Publishing the deduplication bloom filters enables cross-corpus contamination detection. Thematic partitioning supports selective usage based on application requirements. Together with the corpus, we also release a highly scalable preprocessing pipeline with specific considerations for German language. This enables future additions and community contributions to the collection.

7 Conclusion

The *German Commons* is a data collection intended to address a fundamental challenge in open German language model development: the scarcity of large-scale, verifiably licensed training data for German language. By systematically aggregating 154.56 billion tokens of openly licensed German text from institutional sources, this work represents the largest collection of open German text to date and enables the development of language models without the licensing issues prevalent in web-scraped alternative corpora.

The corpus encompasses seven thematic domains, with text from web, political, legal, news, economics, cultural, and scientific sources. We apply systematic quality filtering, deduplication, and PII removal. Through detailed corpus statistics, we show the suitability of the included data for model training, and verify the high quality of the provided text. Every document in the corpus is further tagged with an explicit canonical SPDX license URL, enabling unambiguous downstream use. The *German Commons* thus represent a critical step toward sustainable, ethically compliant development of German language models.

Acknowledgments

This work has been partially funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) under Grants №`01IS24077A`, №`01IS24077B`, and №`01IS24077D`; by the ScaDS.AI Center for Scalable Data Analytics and Artificial Intelligence, funded by the BMFTR and by the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus under Grant №`ScaDS.AI`; and by the OpenWebSearch.eu project, funded by the European Union under Grant №`GA 101070014`.

References

[1] Longpre et al. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *CoRR*, abs/2310.16787. doi:10.48550/ARXIV.2310.16787.

[2] Kilgarriff and Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Comput. Linguistics*, 29, 3, 333–348. doi:10.1162/089120103322711569.

[3] Brin, Motwani, Page, and Winograd. 1998. What can you do with a Web in your Pocket? *IEEE Data Eng. Bull.*, 21, 2, 37–47.

[4] Gebru, Morgenstern, Vecchione, Vaughan, Wallach, III, and Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64, 12, 86–92. doi:10.1145/3458723.

[5] Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21, 140:1–140:67.

[6] Xue, Constant, Roberts, Kale, Al-Rfou, Siddhant, Barua, and Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. ACL, 483–498. doi:10.18653/V1/2021.NAACL-MAIN.41.

[7] Gao et al. 2021. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR*, abs/2101.00027.

[8] Abadji, Suárez, Romary, and Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Proc. of LREC*. European Language Resources Association, 4344–4355.

[9] Laurençon et al. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

[10] Weber et al. 2024. RedPajama: an Open Dataset for Training Large Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

[11] Soldaini et al. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the ACL (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. ACL, 15725–15788. doi:10.18653/V1/2024.ACL-LONG.840.

[12] De Gibert et al. 2024. A New Massive Multilingual Dataset for High-Performance Language Technologies. In *Proc. of LREC/COLING*. ELRA and ICCL, 1116–1128.

[13] Penedo, Kydlícek, Allal, Lozhkov, Mitchell, Raffel, von Werra, and Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

[14] Penedo et al. 2025. FineWeb2: One Pipeline to Scale Them All - Adapting Pre-Training Data Processing to Every Language. *CoRR*, abs/2506.20920. doi:10.48550/ARXIV.2506.20920.

[15] Habernal, Zayed, and Gurevych. 2016. C4Corpus: Multilingual Web-size Corpus with Free License. In *Proc. of LREC*. European Language Resources Association (ELRA).

[16] Il, Bommarito, and Katz. 2025. The KL3M Data Project: Copyright-Clean Training Resources for Large Language Models. *CoRR*, abs/2504.07854. doi:10.48550/ARXIV.2504.07854.

[17] Hong, Hutson, Agnew, Huda, Kohno, and Morgenstern. 2025. A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset. *CoRR*, abs/2506.17185. doi:10.48550/ARXIV.2506.17185.

[18] Longpre et al. 2024. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the ACL: Human Language Technologies (Volume 1: Long Papers)*. ACL, 3245–3276. doi:10.18653/v1/2024.naacl-long.179.

[19] Wang, Zhong, Wang, Zhu, Mi, Wang, Shang, Jiang, and Liu. 2024. Data Management For Training Large Language Models: A Survey. (2024).

[20] Goldhahn, Eckart, and Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proc. of LREC*. European Language Resources Association (ELRA), 759–765.

[21] Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*. European Language Resources Association (ELRA), 2214–2218.

[22] Kocetkov et al. 2023. The Stack: 3 TB of permissively licensed source code. *Trans. Mach. Learn. Res.*, 2023.

[23] Min, Gururangan, Wallace, Shi, Hajishirzi, Smith, and Zettlemoyer. 2024. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

[24] Kandpal et al. 2025. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. *CoRR*, abs/2506.05209. doi:10.48550/ARXIV.2506.05209.

[25] Langlais et al. 2025. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training. *CoRR*, abs/2506.01732. doi:10.48550/ARXIV.2506.01732.

[26] Nolda. Wikipedia-Korpus: Korpusquellen der deutschsprachigen Wikipedia im TEI-Format. Version 20250101. Zenodo. doi:10.5281/zenodo.14748605.

[27] Nolda. Wikivoyage-Korpus: Korpusquellen der deutschen Sprachversion von Wikivoyage im TEI-Format. Version 20250101. Zenodo. doi:10.5281/zenodo.14748553.

[28] Margaretha and Lüngen. 2014. Building Linguistic Corpora from Wikipedia Articles and Discussions. *J. Lang. Technol. Comput. Linguistics*, 29, 2, 59–82.

[29] Schabus, Skowron, and Trapp. 2017. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. ACM, 1241–1244. doi:10.1145/3077136.3080711.

[30] Boenig, Haaf, Nolda, and Wiegand. Reichstagsprotokoll-Korpus. Version v1.0.2. Zenodo. doi:10.5281/zenodo.10225467.

[31] Barbaresi. German Political Speeches Corpus. Version v4.2019. Zenodo. doi:10.5281/zenodo.3611246.

[32] Fobbe. Corpus der Drucksachen des Deutschen Bundestages (CDRS-BT). Version 2021-04-02. Zenodo. doi:10.5281/zenodo.4643066.

[33] Fobbe. Corpus der Plenarprotokolle des Deutschen Bundestages (CPP-BT). Version 2021-02-17. Zenodo. doi:10.5281/zenodo.4542662.

[34] Fobbe. Corpus des Deutschen Bundesrechts (C-DBR). Version 2025-01-07. Zenodo. doi:10.5281/zenodo.14592346.

[35] Ostendorff, Blume, and Ostendorff. 2020. Towards an Open Platform for Legal Information. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*. ACM, 385–388. doi:10.1145/3383583.3398616.

[36] Fobbe. Corpus der Entscheidungen des Bundesfinanzhofs (CE-BFH). Version 2025-01-14. Zenodo. doi:10.5281/zenodo.14622341.

[37] Fobbe and Swalve. Entscheidungen des Bundesgerichtshofs in Strafsachen aus dem 20. Jahrhundert (BGH-Strafsachen-20Jhd). Version 1.0.0. Zenodo. doi:10.5281/zenodo.4540377.

[38] Fobbe. Corpus der Entscheidungen des Bundesgerichtshofs (CE-BGH). Version 2024-09-25. Zenodo. doi:10.5281/zenodo.12814022.

[39] Fobbe. Corpus der Entscheidungen des Bundesverfassungsgerichts (CE-BVerfG). Version 2024-07-24. Zenodo. doi:10.5281/zenodo.12705674.

[40] Fobbe. Corpus der Entscheidungen des Bundespatentgerichts (CE-BPatG). Version 2024-07-09. Zenodo. doi:10.5281/zenodo.10849977.

[41] Fobbe. Corpus der Entscheidungen des Bundesverwaltungsgerichts (CE-BVerwG). Version 2024-03-13. Zenodo. doi:10.5281/zenodo.10809039.

[42] Fobbe. Corpus der amtlichen Entscheidungssammlung des Bundesverfassungsgerichts (C-BVerfGE). Version 2024-03-08. Zenodo. doi:10.5281/zenodo.10783177.

[43] Fobbe. Corpus der Entscheidungen des Bundesarbeitsgerichts (CE-BAG). Version 2020-08-28. Zenodo. doi:10.5281/zenodo.4006645.

[44] Chalkidis, Fergadiotis, and Androutsopoulos. 2021. MultiEURLEX - A multilingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In ACL, 6974–6996. doi:10.18653/V1/2021.EMNLP-MAIN.559.

[45] Boenig and Hug. DiBiLit-Korpus. Version v3.0. Zenodo. doi:10.5281/zenodo.5786725.

[46] Labs. 2021. Digitised Books. c. 1510 - c. 1900. JSONL (OCR derived text + metadata). https://doi.org/10.23636/r7w6-zy15. (2021).

[47] Gaido, Papi, Bentivogli, Brutti, Cettolo, Gretter, Matassoni, Nabih, and Negri. 2024. MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages. In *Proc. of EMNLP*. ACL, 13934–13947.

[48] Labusch and Lehmann. Fulltexts of the Digitized Collections of the Berlin State Library (SBB). Version 1. Staatsbibliothek zu Berlin - Berlin State Library. doi:10.5281/zenodo.7716098.

[49] Nolda. Wikibooks-Korpus: Korpusquellen von deutschsprachigen Wikibooks im TEI-Format. Version 20250101. Zenodo. doi:10.5281/zenodo.14748586.

[50] Hug, Kassung, and Meyer. 2010. Dingler-Online – The Digitized »Polytechnisches Journal« on Goobi Digitization Suite. In *Digital Humanities DH2010. Conference Abstracts*, 311–313.

[51] Priem, Piwowar, and Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *CoRR*, abs/2205.01833. doi:10.48550/ARXIV.2205.01833.

[52] Lee, Cooper, and Grimmelmann. 2024. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version). In *Proceedings of the Symposium on Computer Science and Law, CSLAW 2024, Boston, MA, USA, March 12-13, 2024*. ACM, 48–63. doi:10.1145/3614407.3643696.

[53] Tarkowski. 2025. Data Governance in Open Source AI.

[54] 2025. GROBID. https://github.com/kermitt2/grobid. (2025).

[55] Poznanski, Borchardt, Dunkelberger, Huff, Lin, Rangapur, Wilhelm, Lo, and Soldaini. 2025. olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models. *CoRR*, abs/2502.18443. doi:10.48550/ARXIV.2502.18443.

[56] Speer. 2019. ftfy. Zenodo. (2019). doi:10.5281/zenodo.2591652.

[57] Joulin, Grave, Bojanowski, Douze, Jégou, and Mikolov. 2016. FastText.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

[58] Joulin, Grave, Bojanowski, and Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proc. of EACL*. ACL, 427–431. doi:10.18653/V1/E17-2068.

[59] Radford, Wu, Child, Luan, Amodei, Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1, 8, 9.

[60] Scao et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR*, abs/2211.05100. doi:10.48550/ARXIV.2211.05100.

[61] Rae et al. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *CoRR*, abs/2112.11446.

[62] Nguyen, Nguyen, Lai, Man, Ngo, Dernoncourt, Rossi, and Nguyen. 2024. CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*. ELRA and ICCL, 4226–4237.

[63] Mendels, Peled, Vaisman Levy, Hart, Rosenthal, Lahiani, et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable PII anonymization service for text and images. Microsoft, (2018).

[64] Arnett, Jones, Yamshchikov, and Langlais. 2024. Toxicity of the Commons: Curating Open-Source Pre-Training Data. *arXiv preprint arXiv:2410.22587*.

[65] Guhr, Schumann, Bahrmann, and Böhme. 2020. Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, 1620–1625.

[66] Zhu, Chen, Gao, and Wang. 2024. Evaluating llms at evaluating temporal generalization. *CoRR*, abs/2405.08460. doi:10.48550/ARXIV.2405.08460.

[67] Kanerva, Ledins, Käpyaho, and Ginter. 2025. OCR error post-correction with llms in historical documents: no free lunches. *CoRR*, abs/2502.01205. doi:10.48550/ARXIV.2502.01205.

[68] Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020 [online only]* (CEUR Workshop Proceedings). Vol. 2624. CEUR-WS.org.

[69] Bender and Friedman. 2018. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguistics*, 6, 587–604. doi:10.1162/TACL_A_00041.

A Data Filtering

A.1 Quality Filtering Parameter Values

Parameter	Explanation	Excl. if
Alphabetic Word Ratio	Ratio of whitespace-separated words consisting of only alphabetic characters	> 0.99
Bullet Line Ratio	Ratio of lines starting with • or – characters	> 0.70
Ellipsis Line Ratio	Ratio of lines ending in . . . or . . .	> 0.30
Ellipsis Ratio	Ratio of . . . or . . . substrings occuring to overall whitespace-separated words	> 0.10
Hash Ratio	Ratio of # character occuring to overall whitespace-separated words	> 0.10
Stop-word Count	Number of stopwords in text; for stop words used, see below	< 6.00
Duplicated Line Fraction	Amount of duplicated lines in a document, measured as ratio of lines.	> 0.25
Duplicated Lines Character Fraction	Amount of duplicated lines in a document, measured as ratio of characters.	> 0.15
Duplicated Paragraph Fraction	Amount of duplicated paragraphs in a document, measured as ratio of paragraphs.	> 0.30
Duplicated Paragraph Character Fraction	Amount of duplicated paragraphs in a document, measured as ratio of characters.	> 0.20
Duplicate 5-gram Character Fraction	Text accounted for by duplicated n-grams, measured as ratio of characters.	> 0.39
Duplicate 6-gram Character Fraction		> 0.39
Duplicate 7-gram Character Fraction		> 0.38
Duplicate 8-gram Character Fraction		> 0.38
Duplicate 9-gram Character Fraction		> 0.37
Duplicate 10-gram Character Fraction		> 0.37
Top-2-Gram Character Fraction	Text accounted for by most frequent n-gram, measured as ratio of characters.	> 0.07
Top-3-Gram Character Fraction		> 0.10
Top-4-Gram Character Fraction		> 0.13
Spacing Anomaly Ratio	Ratio of spacing anomalies; missing spaces, excessive spaces, spaced words	> 0.15
Case Anomaly Ratio	Ratio of case anomalies such as random capitalization and mixed case within words	> 0.10
Word Fragment Ratio	Ratio of likely OCR word fragments (1-2 character words excluding common words)	> 0.20
Line Artifact Ratio	Ratio of lines that are likely OCR artifacts (single characters, page numbers)	> 0.25
Special Character Density	Density of unusual unicode characters	> 0.03
Repeated Character Ratio	Ratio of text consisting of repeated character sequences and repeated patterns	> 0.05
Numeric Context Errors	Ratio of numbers inappropriately embedded within words (excluding ordinals)	> 0.08
Avg. Word Length (min)	Average length of whitespace-separated words in characters	< 4.80
Avg. Word Length (max)	Average length of whitespace-separated words in characters	> 7.30
Word Length Standard Deviation (min)	Standard deviation of word lengths in characters	< 1.00
Word Length Standard Deviation (max)	Standard deviation of word lengths in characters	> 5.00
Very Short Words Ratio	Ratio of words with length ≤ 1 character after removing punctuation	> 0.10
Very Long Words Ratio	Ratio of words with length ≥ 15 characters after removing punctuation	> 0.10

A.2 Stopwords

Category	Words
Definite articles	der, die, das, den, dem, des
Indefinite articles	ein, eine, einen, einem, einer
Conjunctions	und, oder, aber
Common Verbs	ist, sind, hat, haben, wird, werden,
Prepositions	von, zu, mit, in, auf, für, bei, nach, vor, über, unter, durch, gegen, ohne, um
Pronouns	ich, du, er, sie, es, wir, ihr, sich, sein, seine, ihrer, ihren, mich, dich
Adverbs	nicht, auch, nur, noch, schon, hier, dort, da, dann, jetzt, heute sehr, mehr, weniger, ganz, gar, etwa
Subordinating Conjunctions	dass, wenn, als, wie
Contractions	an, am, im, ins, zum, zur, vom, beim
Question words	was, wer, wo, wann, warum, wie, welche, welcher
Quantifiers	alle, viele, einige, andere, jede, jeden, jeder
Modal Verbs	kann, könnte, muss, soll, will, würde
Particles	ja, nein, doch, so, also, nun, mal

A.3 PII Generic Replacement Values

PII Category	Generic Replacement	Comment
Credit Card Numbers	4242 4242 4242 4242	VISA testing number; valid but unused
IP Addresses	192.0.2.255	RFC 5737 Test Block 1
Email Addresses	name@beispiel.de	Example domain
Phone Numbers	+49 123 45678910	Invalid number with correct format
IBAN Codes	DE02 1203 0000 0000 2020 51	DKB testing number; valid but unused
URLs	https://www.beispiel.de	Example domain

A.4 Processing Step Statistics

Data Source	Initial†		Filtered				Deduplicated				Final‡			
	# Docs	# Tokens	# Docs		# Tokens		# Docs		# Tokens		# Docs		# Tokens	
Wikipedia	2.940 M	5.981 B	2.931 M	99.68 %	3.037 B	50.77 %	2.930 M	99.66 %	2.949 B	49.30 %	2.930 M	99.66 %	2.949 B	49.30 %
Wikivoyage	0.020 M	0.045 B	0.020 M	99.48 %	0.042 B	94.43 %	0.020 M	99.48 %	0.044 B	99.67 %	0.020 M	99.47 %	0.042 B	94.43 %
Wiki Discussions	8.787 M	2.233 B	8.524 M	96.99 %	1.232 B	55.16 %	8.349 M	95.01 %	1.218 B	54.56 %	8.349 M	95.01 %	1.218 B	54.55 %
Youtube-Commons	3.311 M	15.372 B	2.969 M	89.66 %	14.735 B	95.86 %	2.810 M	84.85 %	14.479 B	94.19 %	2.810 M	84.85 %	14.479 B	94.19 %
One Million Posts Corpus	1.024 M	0.099 B	0.947 M	92.52 %	0.095 B	96.90 %	0.946 M	92.40 %	0.095 B	96.28 %	0.946 M	92.40 %	0.095 B	96.28 %
The Stack	23.089 M	50.004 B	0.512 M	2.22 %	1.465 B	2.93 %	0.478 M	2.07 %	1.370 B	2.74 %	0.421 M	1.83 %	1.105 B	2.22 %
Reichtagsprotokolle	0.001 M	1.115 B	0.001 M	99.05 %	0.763 B	68.37 %	0.001 M	99.05 %	0.703 B	63.08 %	0.001 M	99.05 %	0.703 B	63.08 %
German Political Speeches	0.007 M	0.030 B	0.007 M	99.96 %	0.029 B	99.21 %	0.007 M	99.90 %	0.029 B	99.14 %	0.007 M	99.90 %	0.029 B	99.14 %
C. d. Drucksachen d. dt. BT	0.005 M	0.962 B	0.005 M	95.96 %	0.889 B	92.47 %	0.003 M	59.80 %	0.529 B	54.98 %	0.003 M	59.80 %	0.529 B	54.98 %
C. d. Plenarprotok. d. dt. BT	0.003 M	0.561 B	0.002 M	72.85 %	0.341 B	60.74 %	0.002 M	66.53 %	0.316 B	56.34 %	0.002 M	66.53 %	0.316 B	56.34 %
EuroVoc	5.441 M	49.380 B	0.424 M	7.80 %	3.136 B	6.35 %	0.246 M	4.52 %	1.988 B	4.03 %	0.246 M	4.52 %	1.988 B	4.03 %
Deutsches Bundesrecht	0.007 M	0.114 B	0.004 M	55.31 %	0.001 B	0.93 %	0.003 M	47.18 %	0.001 B	0.88 %	0.003 M	47.18 %	0.001 B	0.88 %
OpenLegalData	0.251 M	2.188 B	0.251 M	99.91 %	2.160 B	98.70 %	0.250 M	99.55 %	1.916 B	87.56 %	0.250 M	99.55 %	1.916 B	87.56 %
C. d. Ents. d. BFH	0.011 M	0.081 B	0.011 M	100.00 %	0.070 B	87.08 %	0.011 M	100.00 %	0.068 B	84.05 %	0.011 M	100.00 %	0.068 B	84.05 %
Ents. d. BGH (20. Jhd.)	0.036 M	0.103 B	0.036 M	99.42 %	0.099 B	96.41 %	0.036 M	99.30 %	0.093 B	90.54 %	0.036 M	99.30 %	0.093 B	90.57 %
C. d. Ents. d. BGH	0.078 M	0.437 B	0.078 M	99.94 %	0.317 B	72.57 %	0.077 M	99.19 %	0.293 B	67.04 %	0.077 M	99.19 %	0.293 B	67.04 %
C. d. Ents. d. BVerfG	0.009 M	0.081 B	0.009 M	99.93 %	0.066 B	81.59 %	0.008 M	89.71 %	0.040 B	48.57 %	0.008 M	89.71 %	0.040 B	48.57 %
C. d. Ents. d. BpatG	0.031 M	0.257 B	0.031 M	99.56 %	0.201 B	78.20 %	0.031 M	99.48 %	0.185 B	72.05 %	0.031 M	99.48 %	0.185 B	72.05 %
C. d. Ents. d. BVerwG	0.027 M	0.182 B	0.027 M	99.99 %	0.143 B	78.37 %	0.027 M	99.94 %	0.123 B	67.73 %	0.027 M	99.94 %	0.123 B	67.73 %
C. d. amtl. E.-S. BVerfG	0.001 M	0.029 B	0.001 M	99.67 %	0.025 B	83.91 %	0.001 M	99.67 %	0.024 B	82.92 %	0.001 M	99.67 %	0.024 B	82.92 %
C. d. Ents. d. BAG	0.006 M	0.062 B	0.006 M	99.98 %	0.060 B	96.76 %	0.006 M	99.98 %	0.048 B	77.71 %	0.006 M	99.98 %	0.048 B	77.71 %
EurLEX	0.065 M	0.205 B	0.065 M	99.91 %	0.201 B	98.05 %	0.065 M	99.90 %	0.201 B	98.04 %	0.065 M	99.90 %	0.201 B	98.05 %
Deutsches Zeitungsportal	10.382 M	60.140 B	8.942 M	86.13 %	47.511 B	79.00 %	8.076 M	77.79 %	43.866 B	72.94 %	8.076 M	77.79 %	43.871 B	72.95 %
Europeana Newspapers	4.125 M	26.904 B	3.296 M	79.90 %	20.727 B	77.04 %	3.256 M	78.94 %	20.676 B	76.85 %	3.256 M	78.94 %	20.684 B	76.86 %
Wikinews	0.042 M	0.016 B	0.027 M	64.73 %	0.015 B	91.51 %	0.023 M	55.96 %	0.014 B	87.59 %	0.023 M	55.95 %	0.014 B	87.59 %
Anno	2.220 M	10.005 B	1.910 M	86.04 %	8.104 B	81.00 %	1.910 M	86.04 %	8.103 B	80.99 %	1.910 M	86.04 %	8.104 B	81.00 %
TEDEUTenders	0.224 M	0.989 B	0.061 M	27.11 %	0.197 B	19.97 %	0.057 M	25.54 %	0.111 B	11.19 %	0.057 M	25.54 %	0.111 B	11.19 %
DiBiLit-Korpus	0.002 M	0.291 B	0.002 M	99.14 %	0.224 B	77.10 %	0.002 M	99.04 %	0.216 B	74.35 %	0.002 M	99.04 %	0.216 B	74.35 %
DiBiPhil-Korpus	0.000 M	0.033 B	0.000 M	99.99 %	0.033 B	98.86 %	0.000 M	99.99 %	0.032 B	97.72 %	0.000 M	99.99 %	0.032 B	97.72 %
Wikisource	0.576 M	0.701 B	0.482 M	83.58 %	0.574 B	81.82 %	0.241 M	41.77 %	0.348 B	49.59 %	0.241 M	41.77 %	0.348 B	49.59 %
German-PD	0.132 M	62.003 B	0.124 M	94.22 %	51.127 B	82.46 %	0.124 M	93.98 %	49.322 B	79.55 %	0.124 M	93.74 %	49.333 B	79.34 %
BLBooks	0.004 M	2.133 B	0.004 M	95.96 %	1.011 B	47.39 %	0.004 M	93.82 %	1.011 B	47.39 %	0.004 M	93.81 %	1.012 B	47.39 %
SBB Fulltexts	4.988 M	4.367 B	3.162 M	63.39 %	3.218 B	73.68 %	3.127 M	62.69 %	3.182 B	72.85 %	3.127 M	62.69 %	3.182 B	72.85 %
Wikiquote	0.011 M	0.007 B	0.009 M	84.01 %	0.007 B	94.99 %	0.009 M	81.43 %	0.007 B	91.50 %	0.009 M	81.43 %	0.007 B	91.54 %
MOSEL	2.984 M	0.388 B	2.618 M	87.76 %	0.360 B	92.93 %	2.606 M	87.32 %	0.359 B	92.48 %	2.606 M	87.32 %	0.359 B	92.48 %
Dig. d. Polytechn. Journals	0.000 M	0.301 B	0.000 M	100.00 %	0.185 B	61.37 %	0.000 M	100.00 %	0.180 B	59.93 %	0.000 M	100.00 %	0.180 B	59.93 %
Wikibooks	0.030 M	0.083 B	0.028 M	91.86 %	0.057 B	68.40 %	0.027 M	90.96 %	0.050 B	60.73 %	0.027 M	90.96 %	0.050 B	60.73 %
DOAB	0.002 M	0.319 B	0.002 M	97.53 %	0.300 B	94.00 %	0.002 M	92.20 %	0.167 B	52.35 %	0.002 M	92.20 %	0.167 B	52.35 %
arXiv	0.461 M	5.746 B	0.000 M	0.02 %	0.001 B	0.02 %	0.000 M	0.00 %	0.000 B	0.00 %	0.000 M	0.00 %	0.000 B	0.00 %
Wikiversity	0.025 M	0.035 B	0.020 M	78.61 %	0.030 B	87.80 %	0.016 M	65.50 %	0.028 B	80.12 %	0.016 M	65.49 %	0.028 B	80.11 %
OpenALEX	0.116 M	0.646 B	0.049 M	42.59 %	0.478 B	73.89 %	0.048 M	41.25 %	0.414 B	64.99 %	0.048 M	41.25 %	0.414 B	63.99 %
Total	71.473 M	304.629 B	37.594 M	52.60 %	163.266 B	53.59 %	35.835 M	50.14 %	154.799 B	50.81 %	35.778 M	50.06 %	154.558 B	50.73 %

Percentages indicate remaining of initial. † After filtering using the source datasets' metadata, if available. ‡ Includes PII replacement and final license filtering.

Datasheet: German Commons

Motivation

Why was the dataset created? German Commons addresses the critical gap in large-scale open German text for language model training. Existing German corpora either lack explicit licensing, contain web-scraped content of uncertain provenance, or provide insufficient scale.

Has the dataset been used already? This represents the initial release of German Commons. No external usage has occurred prior to publication.

What (other) tasks could the dataset be used for? Beyond language model pretraining, German Commons supports all German NLP research requiring clean, license-compliant text, multilingual model development, or linguistic analysis of German text across domains. The diverse domain coverage (legal, cultural, scientific, etc.) further enables domain-specific model development and cross-domain evaluation studies.

Who funded the creation of the dataset? Dataset compilation was supported by German and European research grants: German Federal Ministry of Research, Technology, and Space (BMFTR) under Grants № 01IS24077A, № 01IS24077B, and № 01IS24077D, by the ScaDS.AI Center for Scalable Data Analytics and Artificial Intelligence, funded by the BMFTR and by the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus under Grant № ScaDS.AI, and by the OpenWeb-Search.eu project, funded by the European Union under Grant № GA 101070014. Constituent datasets originate primarily from state-funded institutions across Germany and Austria.

Dataset Composition

What are the instances? Each instance represents a single German-language document with associated metadata and licensing information.

How many instances are there in total? The dataset contains 35,778,211 documents comprising 154,558,196,961 GPT-2 tokens.

What data does each instance consist of? Each instance includes: a unique identifier for source cross-referencing, source dataset name, quality-filtered and paragraph-deduplicated raw text, canonical SPDX license URL, thematic domain key, GPT-2 token count, perplexity score from a German Wikipedia KenLM model, and a OCR quality score.

Is there a label or target associated with each instance? No supervised labels exist. However, each instance contains metadata labels for thematic domain classification, licensing information, and document length statistics.

Is any information missing from individual instances? Paragraph-level deduplication may alter texts from their original form. Personally identifiable information has been systematically removed.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The dataset represents a filtered subset of source collections. Filtering removes OCR errors, extraction artifacts, and low-quality or duplicated content, creating a curated selection.

Are there recommended data splits? No predefined splits are provided. All data is intended for pretraining.

Are there any errors, sources of noise, or redundancies in the dataset? Despite quality filtering and deduplication, residual issues may remain: (1) cross-corpus text duplicates from overlapping sources, and (2) extraction artifacts from OCR and PDF-to-text processing.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? The dataset is self-contained and centrally downloadable. The Source dataset references provided enable reproducible reconstruction.

Collection Process

What mechanisms or procedures were used to collect the data? Data collection employed multiple automated procedures: (1) direct download from institutional repositories and open platforms, (2) programmatic crawling via APIs where available, and (3) automated text extraction from PDF and other document formats using specialized libraries. Then, the open source processing pipelines were applied for quality filtering and deduplication all sources. Validation occurred through manual inspection of sample outputs, cross-verification against source repositories, and automated consistency checks.

How was the data associated with each instance acquired? All text data represents directly observable content from original sources; no inference or derivation occurred. Metadata (licensing, thematic classification, source attribution) was extracted directly from source repository information or explicitly provided by institutional datasets. Where PDF extraction was required, raw text underwent validation against source documents to verify accuracy.

If the dataset is a sample from a larger set, what was the sampling strategy? Sampling was deterministic based on explicit criteria: (1) German language content as per automated classification (2) explicit open licensing, (3) quality thresholds, and (4) institutional source verification. No probabilistic sampling occurred; all content meeting inclusion criteria was retained after deduplication.

Who was involved in the data collection process and how were they compensated? Data collection was conducted by the author team using automated systems. No crowdworkers, contractors, or external annotators were employed. All processing occurred through programmatic methods without manual content creation or labeling requiring compensation.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances? Collection occurred between January and August 2025, using source dataset versions available through August 31st, 2025. The underlying content creation spans multiple centuries, representing a temporal range that significantly predates and extends beyond the collection timeframe.

Data Preprocessing

Was any preprocessing/cleaning/labeling of the data done? Comprehensive preprocessing included: (1) text extraction from PDFs and OCR sources with encoding normalization, (2) language detection and filtering for German content, and (3) quality filtering targeting digitization artifacts and extraction errors, (4) paragraph-level deduplication using content hashing, (5) systematic PII removal, (6) format standardization across all source types. Thematic domain classification was applied based on source dataset.

Was the raw data saved in addition to the preprocessed/cleaned/labeled data? Raw data is not provided since all constituent source datasets remain publicly accessible through their original repositories.

Is the software used to preprocess/clean/label the instances available? All preprocessing software is open source and available at https://github.com/coral-nlp/german-commons and https://github.com/coral-nlp/llmdata , ensuring complete reproducibility of the dataset.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes. The procedure successfully addresses the identified gap by: (1) providing the largest collection to-date of openly licensed German text, (2) enabling open German language model development without licensing uncertainties, and (2) establishing reproducible methodology for future dataset construction. This directly fulfills the stated motivation of creating license-compliant, large-scale German training data.

How will the dataset be distributed? The dataset is distributed as Parquet files through multiple public repositories for redundancy. Primary distribution occurs via Hugging Face Hub at https://huggingface.co/datasets/coral-nlp/german-commons.

When will the dataset be released/first distributed? What license (if any) is it distributed under? Public release occurred on 2025/10/14. Dataset metadata and compilation are licensed under ODC-BY 1.0. Individual document texts retain their original licenses as specified in each instance's SPDX URL field, creating a heterogeneous but fully documented licensing structure.

Are there any copyrights on the data? Yes. Each document retains copyright under its original creator or institutional provider, governed by the specific license indicated in the instance metadata. The compilation itself does not claim additional copyright over constituent texts.

Are there any fees or access/export restrictions? The dataset is freely accessible without fees or registration requirements. However, users must comply with individual document licenses, which may include attribution requirements or share-alike provisions. Commercial use is permitted by all constituent licenses.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset is maintained by the authors of this report.

Will the dataset be updated? If so, how often and by whom? Updates may occur when significant new German open-source collections become available. The original authors will coordinate updates, with community contributions welcomed through the open-source pipeline.

How will updates be communicated? Updates will be announced through: (1) versioned releases on hosting platforms with detailed changelogs, (2) academic publication updates when substantial changes occur.

If the dataset becomes obsolete how will this be communicated? Obsolescence will be communicated through deprecation notices on all hosting platforms.

Is there a repository to link to any/all papers/systems that use this dataset? No centralized usage repository will be maintained. Usage tracking occurs through standard academic citation of the dataset paper. Users are encouraged to cite the dataset publication when reporting results or building derivative works.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? The open-source `llmdata` pipeline enables community extensions through standardized data ingestion protocols for new sources and automated quality assessment and deduplication using established filtering criteria. Community contributions undergo review by the maintenance team.

Ethical Considerations

Were any ethical review processes conducted? No formal institutional review board process was conducted. The dataset relies exclusively on pre-existing, publicly available, and explicitly licensed materials from established institutional sources. Data processing incorporated ethical considerations including systematic PII removal and exclusion of sources lacking clear licensing frameworks.

Does the dataset contain data that might be considered confidential? No. All included content derives from explicitly open-licensed institutional sources.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? Potentially yes. The dataset spans centuries of German text documents, which may include historical perspectives, political viewpoints, or language that could be considered offensive by contemporary standards. The scale and temporal range make comprehensive content moderation infeasible. Users should exercise appropriate caution.

Does the dataset relate to people? The dataset may contain publicly available information relating to individuals in various contexts including historical documents, biographical information, academic citations, and government records.