SciLaD: A Large-Scale, Transparent, Reproducible Dataset for Natural Scientific Language Processing

Luca Foppiano[3,1] Sotaro Takeshita[5] Pedro Ortiz Suarez[4]
Ekaterina Borisova[1] Raia Abu Ahmad[1] Malte Ostendorff[4]
Fabio Barth[1] Julian Moreno-Schneider[1] Georg Rehm[1,2]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany
[2]Humboldt-Universität zu Berlin, Germany [3]ScienciaLAB, Portugal
[4]Common Crawl Foundation, USA [5]University of Mannheim, Germany

Abstract

SciLaD is a novel, large-scale dataset of scientific language constructed entirely using open-source frameworks and publicly available data sources. It comprises a curated English split containing over 10 million scientific publications and a multilingual, unfiltered TEI XML split including more than 35 million publications. We also publish the extensible pipeline for generating SciLaD. The dataset construction and processing workflow demonstrates how open-source tools can enable large-scale, scientific data curation while maintaining high data quality. Finally, we pre-train a RoBERTa model on our dataset and evaluate it across a comprehensive set of benchmarks, achieving performance comparable to other scientific language models of similar size, validating the quality and utility of SciLaD. We publish the dataset and evaluation pipeline to promote reproducibility, transparency, and further research in natural scientific language processing and understanding, including scholarly document processing.

Keywords: Dataset, scientific, natural scientific language processing, LLM, encoder, text mining

1. Introduction

High-quality accessible datasets are a cornerstone of progress in natural language processing (NLP). In recent years, the availability of large-scale open corpora has enabled significant advances across tasks such as machine translation, summarization, and question answering (Raffel et al., 2023; Devlin et al., 2019). However, while much of this progress has been driven by general-domain data, scientific language remains comparatively underrepresented in open datasets, limiting the reproducibility and transparency of research in specialized domains (Lo et al., 2020; Beltagy et al., 2019).

Scientific text exhibits unique linguistic and structural properties – precise terminology, formal discourse organization, and data-centered argumentation – that differ markedly from everyday language. Creating large, openly available, and well-structured corpora of scientific publications is, thus, essential for training and evaluating language models in the scientific domain (Fisas et al., 2015). Resources such as S2ORC (Lo et al., 2020) and CORE (Knoth et al., 2023) have demonstrated the value of openly licensed scientific data; however, few offer fully transparent, reproducible pipelines built exclusively on open-source frameworks.

We present SciLaD (Scientific Language Dataset), a large-scale dataset of scientific publications created entirely using open-source tools, ensuring reproducibility, transparency, and accessibility. The dataset is harvested from Open Access (OA) scientific articles and represented in TEI XML (Text Encoding Initiative), preserving document structure and metadata, which is essential for downstream applications such as citation graph analysis, bibliometric research, and retrieval-based tasks.

To demonstrate the utility and quality of SciLaD, we train and evaluate a RoBERTa-base (Liu et al., 2019) encoder model. The model and its tokenizer are pre-trained from scratch using our data, SciLaD-M (CUSTOM). For comparison, we also train a variation, which uses the original RoBERTa tokenizer, SciLaD-M (RoBERTa). The model SciLaD-M (CUSTOM) achieves results comparable to other domain-specific encoders such as SciBERT (Beltagy et al., 2019), validating the effectiveness of the dataset as a foundation for scientific NLP. Furthermore, we introduce a reproducible evaluation pipeline to benchmark encoder and encoder-decoder architectures on multiple scientific tasks, providing a framework for systematic comparison of models trained on scientific corpora.

Our main contributions are:

- A large-scale multilingual corpus of over 35 million scientific articles sourced from Unpaywall[1], structured in TEI XML format[2], suitable for various upstream and downstream tasks, including bibliographic, citation graph analysis (109 million citations), and citation graph

[1]https://unpaywall.org
[2]https://tei-c.org

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7606–7618
11–16 May 2026.

or characterization training, as well as classic NLP tasks in the scientific domain. We also created a filtered English split of 10 million text documents for pre-training.

- An open source pre-processing pipeline for the dataset using frameworks such as Grobid (GROBID contributors, 2008 - 2025) and Datatrove (Penedo et al., 2024b) as well as an evaluation pipeline to fine-tune and evaluate encoder- or encoder-decoder-based models on common scientific NLP tasks.

- Encoder-based scientific language models based on RoBERTa-base architecture (Liu et al., 2019) (110M parameters) and pre-trained from scratch with a clean, plain text dataset of over 10 million scientific publications in English. We evaluate the produced models on multiple benchmarks and compare them with a set of baseline models.

The resources shared through this article are described in Section 6 and will be made publicly available upon acceptance.

2. Dataset Construction

The dataset was constructed using the open-source SciLaD pipeline, which automates the retrieval, filtering, and preprocessing of scientific publications. The pipeline integrates multiple components for metadata extraction, text cleaning, and language identification, ensuring consistent data quality across languages and scientific domains. By relying exclusively on open-access sources and transparent workflows, we enable reproducibility and facilitate community-driven extensions of the dataset. In the following, we provide brief descriptions of the pipeline architecture and each stage of the data construction process, from harvesting to preprocessing and evaluation.

2.1. Processing pipeline

The pipeline is composed of three steps (Figure 1). First, articles are collected and downloaded following the Unpaywall snapshot and complemented with additional formats, including JATS (Journal Article Tag Suite) XML and LaTeX from other sources, e. g., arXiv and PubMed Central (PMC). Unpaywall is a database that indexes over 50 million open-access scholarly articles by aggregating metadata from trusted repositories, journals, and archives. It provides DOI-based access to legally available full texts, which we used as the primary data source for collecting multilingual documents across scientific domains. In the second step, these articles were then transformed into a standardized, structured format (TEI XML). In the third step, the data was converted into plain text. After extracting the plain text, we applied data quality filtering and deduplication.

2.1.1. Harvesting

The main challenge in accessing scientific publications is that they are typically published behind paywalls, restricting automated access. However, a significant number of publications are now available in OA (Piwowar et al., 2018), making it possible to access these resources legally while complying with constraints related to text and data mining and fair use. However, even for OA publications, the dominant scientific publishers often use mechanisms to limit access (access limitation and strict policies), especially via programmatic means, which implies the need for robust harvesting methods.

As illustrated in Figure 1, we rely on Unpaywall for accessing PDF files (Piwowar et al., 2018; Else, 2018; Chawla, 2017). We used the full snapshot produced on 14 June 2023, a total of 38.9 million entries with a non-empty `url_for_pdf` field. These entries, however, are not only documents but also other "components" that represent elements of documents (figures, tables, etc.), see Table 1. After removing these entries and downloading the PDF files corresponding to each URL, we obtained 31.6 million documents. These documents are, at this stage, neither filtered nor deduplicated. Although the majority (77.1%) of the data is in English, there are still some texts in other languages (see Figure 2a). To facilitate access to LaTeX and JATS XML we mirrored arXiv[3] resources (around 2 million OA articles with LaTeX sources), PMC[4] full texts (around 4 million articles in JATS XML) and the PLOS[5] (Seiver et al., 2018) XML full text collections (around 300,000 articles), all available publicly online, on different AWS (Amazon Web Service) S3 buckets.

Description	Count
Total Unpaywall entries (with non-empty `url_for_pdf`)	38,931,157
Total entries (excluding type "component")	37,451,475
Total harvested PDF documents	31,653,759

Table 1: Overview of the PDF files collected from the Unpaywall snapshot. The entries with a non-empty URL-to-PDF are retained, then filtered to remove "components" and further reduced by about 6 million documents that fail to download due to incomplete URLs, limiting landing pages, etc.

The harvesting and processing of these files has

[3] https://info.arxiv.org/help/bulk_data_s3.html

[4] http://pmc.ncbi.nlm.nih.gov

[5] https://github.com/PLOS/allofplos

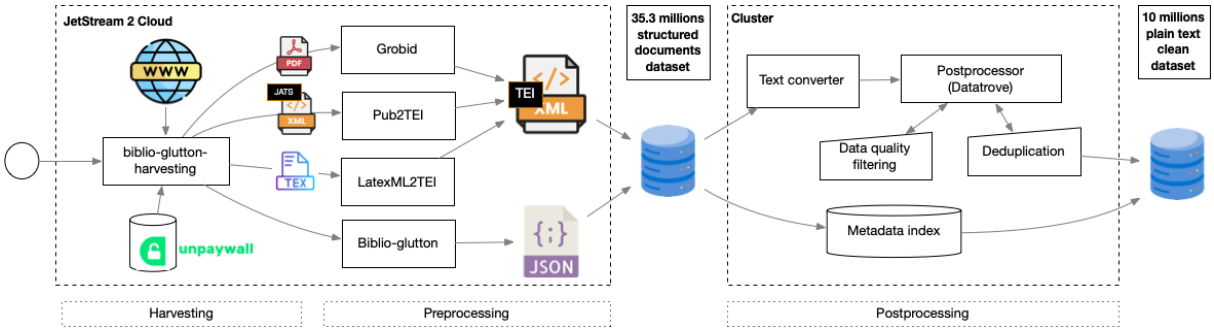


Figure 1: The SciLaD pipeline. The data is harvested by accessing direct links to the source files and is transformed into TEI XML. The structured format's scientific data is streamed to text, while its bibliographic data are aggregated in a metadata index. Finally, the text is filtered and deduplicated.

been performed on the JetStream 2 HPC environment. We divided the Unpaywall snapshot into 3,000 partitions and used 24 virtual machine instances in parallel to harvest and process the set of documents. Each instance had 8 CPUs, 32 MB of RAM, and one GPU with 8 GB of VRAM.

2.1.2. Preprocessing

The biggest challenge in extracting a structured representation of published content is the immense variety of formats and presentation layers in scientific publications, which makes it very difficult to reliably extract and normalize content into a single structured format. The extraction of content from arbitrary PDF files is a well-known challenge (Westergaard et al., 2017). As mentioned, we normalize the three heterogeneous data types (PDFs, JATS XML, and LaTeX) to a standardized, structured format, i. e., TEI XML, which we then transform into a text format in the preprocessing step.

Given the extreme volume of publications to process, we did not use Visual Language Models (VLM) to extract the text. Using VLMs may yield higher accuracy compared to other LLM-based approaches, however, the processing is too slow and too expensive to scale beyond 500 million pages (Blecher et al., 2023). For example, Nougat, a Visual Transformer that focuses on scientific papers, requires an A10 with 24GB VRAM to process six pages per batch, with a processing time of 19.5 seconds per batch, i. e., 3.25 seconds per page (Blecher et al., 2023). Other API-based LLMs, such as ChatGPT, also offer the possibility of extracting text from PDFs, but their closed-source nature makes the reproducibility of our work impossible, and the costs for extraction are much too high.

We, therefore, decided to use an open source software that is not based on a VLM: Grobid (GROBID contributors, 2008 - 2025) relies on a set of custom Recurrent Neural Network (RNN) and Conditional Random Field (CRF) models that work well on commodity hardware. Grobid can process, on average, around 120 pages per second on a single commodity server, which is approximately 400 times faster than Nougat.

It is considered that VLMs offer more reliable extraction of formulas and figures. However, (Poznanski et al., 2025) shows that, for end-to-end tasks, results with Grobid-extracted full texts (based on an old version of Grobid from 2019) are comparable to those of the recent state-of-the-art VLM olmOCR. As Grobid extracts the text layer of a PDF file, it avoids errors coming from the Optical Character Recognition (OCR) processing. Nowadays, the large majority of scientific PDF files are born digital, i. e., they have a reliable text layer. With Grobid (GROBID contributors, 2008 - 2025), we processed roughly 31 million scientific articles (Table 2).

Although LaTeX and JATS XML are already structured formats, we also transformed them into TEI XML. For converting JATS XML to TEI XML, we used Pub2TEI, a framework that standardizes heterogeneous XML representations from different scientific publishers into TEI XML (Patrice Lopez, 2014 - 2025). It relies on a comprehensive set of XSLT stylesheets that map publisher-specific metadata and structural elements – such as bibliographic records, abstracts, citations, and full texts – into a consistent schema. For the LaTeX to TEI XML conversion, we used LaTeXML, a software system designed to convert LaTeX documents into structured XML representations suitable for further processing and web publication (Ginev et al., 2014). It emulates the behavior of the TeX engine to interpret document content and structure, producing an intermediate XML format that preserves both textual and mathematical semantics. This format can then be transformed, among others, into TEI XML.

After harvesting and preprocessing the dataset, we obtained 35.3 million publications in TEI XML (Table 2). The dataset's original format was unequally distributed, with a predominance of PDF with 87% (31 million), followed by JATS XML and LaTeX, 8.5% (3 million) and 3.6% (1.2 million), re-

Description	Count
Grobid (PDF files from Unpaywall)	31,062,728
Pub2TEI (JATS XML files from PMC/PLoS)	3,008,154
LaTeXML (LaTeX files from arXiv)	1,256,667
Total	35,342,549

Table 2: Distribution of structured documents by source type. The Grobid row originates from "Total harvested documents" in Table 1 where about 500,000 documents have not been processed by Grobid, likely because they are corrupted, do not contain any, or too much text (> 1.5 M tokens), etc.

spectively (see Table 2). English was the most common language, followed by Japanese, Portuguese, and Spanish (Figure 2a). The TEI XML provides a structured format with 109 million nodes (references cited) and 1.51 billion edges linking citations with their contextualized sentences. Furthermore, we obtained approximately 19 million unique keywords assigned by authors and published in the headers of the articles.

Figure 2b illustrates the distribution of licenses, with particular attention to OA licenses. Although Creative Commons (CC) licenses account for 37.9%, the majority of articles have unidentified licenses (54%). From our study, it also emerges that reusable content accounts for 32.4% of the harvested OA documents. We want to point out that licensing is a gray area that is often overlooked; not all OA content can be redistributed, and the information is, in most cases, challenging to find because it is scattered around several places: articles, landing pages, and the respective publisher's or journal's policies on data sharing. At this stage the dataset consists of 35.3 million publications in TEI XML (Table 2), including PDF documents and data transformed from JATS XML and LaTeX.

2.1.3. Postprocessing

We prepared the harvested data for pre-training by converting the documents into plain text, selecting only English documents, applying text quality filtering, and deduplication steps.

TEI XML to Text The TEI XML to Text conversion was performed using a configurable streaming process to retain only minimal structure. During text processing, all paragraphs were collected, including captions, while formulas, equations, and table bodies were excluded. Equations and table content strongly depended on how the PDF file was created, making it challenging to assess proper quality extraction, avoiding a decrease of text quality. Within each paragraph, detected sentences were joined together, with a single space when necessary. References were maintained in the text through their original callout formats. During conversion, we computed weighted language detection scores aggregated from line-by-line results (Abadji et al., 2022) using an optimized version of fasttext (Joulin et al., 2016; Burchell et al., 2023).

Filtering The filtering was divided into two parts. We first filtered all non-English text from the corpus and then implemented a pipeline with Datatrove (Penedo et al., 2024b), a scalable open-source framework for data cleaning and preparation. Initial filtering was performed using the language-detection scores. All documents with a language detection confidence score below 80% were excluded. This allowed the removal of all documents whose output was deemed unreliable due to character encoding, encryption and incorrect OCR results performed in the past. Furthermore, we applied the Gopher language quality filtering (Rae et al., 2021) with the same parameters as described in the original work. We did not use any other filters provided by Datatrove, such as C4 (Raffel et al., 2023) or FineWeb (Penedo et al., 2024a), because they are designed for web data, which is expected to be noisier than text from scientific articles.

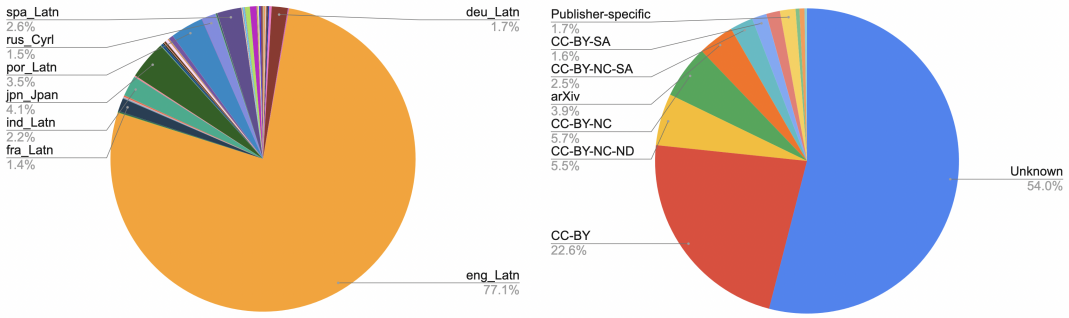
Deduplication We applied MinHash deduplication on the document level to each individual filtered dump. We utilized the deduplication pipeline from Datatrove (Penedo et al., 2024b), which was also used for FineWeb (Penedo et al., 2024a). For this, we generated 5-grams using a word tokenizer (Bird et al., 2009) and computed MinHashes using 112 hash functions, the same setup was used for the FineWeb deduplication. We also targeted documents that are at least 75% similar.

2.2. Corpus Overview

The final corpus obtained for English consisted of 68.7 billion tokens, comprising 10,999,210 documents (with an average of approximately 6,000 tokens per document). Filtering and deduplication reduced the corpus of about 20 million English articles by 50%. The pipeline steps, including data harvesting and TEI XML to text conversion, are shared on GitHub, as discussed in Section 6.

3. Experimental Setup

We pre-trained an encoder-based model to showcase our dataset's utility for model development. To evaluate the pre-trained language model, we developed a comprehensive evaluation pipeline that is easy to use and publicly available. Unlike current benchmarking pipelines, such as Eval-Harness (Gao et al., 2024), our pipeline focuses on scientific



(a) Language distribution

(b) License distribution

Figure 2: Distribution of licenses and languages of the harvested dataset.

benchmarks where models need to be fine-tuned for downstream task.

Our goal is to demonstrate that the corpus is a valuable asset for the research and open-source community, which is why we decided not to train a generative model but rather to use a model architecture comparable to established scientific language models (Beltagy et al., 2019). This decision was also supported by a lower economic footprint and reduced costs for reproducibility.

In this section, we first describe the pre-training process and then outline the evaluation pipeline and benchmarking process.

3.1. Language Model Pre-Training

Language Models To assess the quality of our dataset, we pre-train the model from scratch (with random weights initialization) instead of starting from an existing checkpoint. We obtain SciLaD-M (custom) by training both the model and the tokenizer using our data. The resulting custom tokenizer has the same vocabulary size as the original RoBERTa tokenizer, of 50265 tokens. To make the comparison fairer and transparent, we also train a version that uses the original RoBERTa tokenizer: SciLaD-M (RoBERTa).

We use the same model architecture and training objective as RoBERTa (Liu et al., 2019), namely the `base` architecture[6], composed of a stack of 12 Transformer layers with self-attention, trained on a masked language modeling (MLM) objective.

We follow the hyperparameters described in the original RoBERTa configuration, and we set the batch size ($= 256$), with 2 gradient accumulation steps, which makes the effective batch size ($= 1024$). We trained the model on two H200 with 80Gb VRAM GPUs. We split our dataset into 80/10/10 (%) for training, validation, and held-out test splits, respectively.

[6] https://huggingface.co/FacebookAI/roberta-base

3.2. Evaluation Pipeline

We provide an open and user-friendly evaluation pipeline that allows researchers to fine-tune and evaluate language models on our dataset. The pipeline is built entirely using open-source components and supports a wide range of encoder, decoder, and sequence-to-sequence architectures (e. g., `BERT` (Devlin et al., 2019), `T5` (Raffel et al., 2023), `mBART` (Liu et al., 2020)). It can be easily deployed online or locally, ensuring reproducibility and comparability across experiments.

The evaluation interface standardizes preprocessing, tokenization, and scoring across tasks, producing unified metrics that facilitate benchmarking multilingual scientific language understanding.

Benchmarks To evaluate the models, we followed the benchmarks used by Beltagy et al. (2019), which focus on 1. Named Entity Recognition (NER); 2. Participants, Interventions, Comparisons, and Outcomes (PICO) Extraction; 3. Dependency Parsing (DEP); 4. Relation Classification (REL); and 5. Text Classification (CLS). We released an evaluation framework (see Section 6) that fine-tunes Hugging Face models on all reported benchmarks, outputting their evaluation results.

NER is evaluated using four different benchmarks, three of which tackle the biomedical field, while the last is focused on computer science (CS). BC5CDR (Li et al., 2016) is a widely-used benchmark for evaluating biomedical NER and relation extraction (RE), consisting of 1,500 PMC abstracts manually annotated for chemical and disease mentions as well as their interactions. Similarly, JNLPBA (Collier et al., 2004) evaluates biomedical NER, consisting of 2,000 MEDLINE abstracts annotated with five entity types: protein, DNA, RNA, cell type, and cell line. NCBI-disease (Doğan et al., 2014) focuses on disease entities in PMC abstracts, with 6,892 disease mentions mapped to 790 unique disease concepts. Finally, SciERC (Luan et al., 2018) is a benchmark for joint

scientific extraction tasks, comprising 500 scientific abstracts from AI conference proceedings. It combines NER with RE and co-reference resolution, and includes six entity types such as Task, Method, and Metric, with seven relation types such as Compare and Part-of.

PICO is a NER-like task that focuses on clinical trial papers. To evaluate it, we use the EBM-NLP benchmark (Nye et al., 2018), consisting of 5,000 annotated abstracts. For DEP evaluation, we use GENIA (Kim et al., 2003), a semantically annotated corpus of 2000 abstracts from biological articles. REL is evaluated on two datasets, ChemProt (Kringelum et al., 2016), a benchmark with chemical-protein-disease annotations, and SciERC. Finally, CLS is evaluated using ACL-ARC (Jurgens et al., 2018) and SciCite (Cohan et al., 2019) for citation intent classification from NLP, CS, and biomedical articles. As well as the field of research classification using the Microsoft Academic Graph's Paper Field annotations (Sinha et al., 2015), spanning over seven academic fields.

Evaluation Metrics We compute F1 scores for NER (macro average), PICO, REL, and CLS (macro and micro average). To accommodate for the benchmark annotations, NER results are measured on the span-level, PICO on the token-level, while REL and CLS are both reported on the sentence-level. For DEP, we report both the unlabeled attachment score (UAS) and the labeled attachment score (LAS). The scores are averaged over three runs with three random seeds.

4. Results and Discussion

Before discussing the results, it is important to note that we fine-tuned all baseline models, SciBERT (Beltagy et al., 2019), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), on each task using the optimal hyperparameters provided in their respective publications. We evaluated our two variants SciLaD-M (custom) and SciLaD-M (RoBERTa). Table 3 presents the results across the five task categories and the eleven datasets.

Across all evaluated tasks, our models perform comparably to established domain-specific models, with an overall average score of **74.91** for SciLaD-M (custom) and **73.55** for SciLaD-M (RoBERTa), slightly improving on the performance of SciBERT (**71.19**). This demonstrates that our dataset enables robust representation learning on scientific text, even when trained exclusively on open-access data.

Named Entity Recognition (NER). On the NER tasks, SciLaD-M (custom) achieves the highest F1 score on the BC5CDR dataset (**97.20**), surpassing all baselines, performing on par with SciBERT on JNLPBA, and with SciLaD-M (RoBERTa) on NCBI-disease. Although performance drops on the SciERC dataset – where SciBERT remains the strongest – our models still outperform general-domain baselines such as BERT and RoBERTa, indicating that domain-specific data and tokenization significantly benefit biomedical and scientific entity recognition.

PICO Extraction. For the EBM-NLP dataset, our SciLaD-M (RoBERTa) model achieves the best overall performance (**78.47**), slightly outperforming SciBERT (**78.14**) and demonstrating that our corpus provides competitive representations for structured medical text extraction.

Relation Extraction (REL). In the relation extraction tasks, SciLaD-M (custom) achieves the best score on both ChemProt (**83.83**) and SciERC (**81.24**), surpassing all other baselines. This suggests that our scientific corpus captures relational semantics effectively, even across heterogeneous domains.

Text Classification (CLS). In classification tasks such as CitationIntent, MAG, and SciCite, our models perform competitively. SciLaD-M (custom) obtains the best results on CitationIntent (**64.46**), outperforming all baselines by a notable margin, while performance on MAG and SciCite remains close to that of SciBERT and RoBERTa. These findings indicate that our models generalize well across citation and intent classification tasks.

Dependency Parsing (DEP). For dependency parsing on GENIA, evaluation metrics are UAS and LAS. UAS does not consider the semantic relation (e. g., Subj) used to label the attachment between the head and child, while LAS requires a correct label for each attachment. Both of our variants outperform all baselines, highlighting the potential of our dataset to enhance syntactic understanding in scientific text.

Overall Analysis. In aggregate, the SciLaD-M (custom) and SciLaD-M (RoBERTa) models perform on par with, or better than, the baselines. These results validate the strength of our dataset and demonstrate that open-access scientific corpora can yield competitive language models without relying on proprietary, restricted, or even illegally acquired data sources. Our findings underline that reproducibility and openness do not come at the expense of performance, and that the proposed dataset provides a strong foundation for advancing research in scientific language processing.

	Dataset	SciLaD-M (ours) Custom	SciLaD-M (ours) RoBERTa	SciBERT	BERT	RoBERTa	ModernBERT
NER	BC5CDR‡	**97.20±0.10**	96.54±0.33	95.07±1.06	85.83±0.17	88.07±2.03	94.97±0.24
	JNLPBA‡	93.89±0.28	93.50±0.15	**94.19±0.47**	92.21±1.30	92.11±0.47	91.51±0.31
	NCBI-disease‡	91.48±0.21	**91.63±0.07**	87.09±3.20	73.65±9.05	88.81±0.67	87.74±0.83
	SciERC‡	41.03±0.45	42.49±1.33	**54.39±5.74**	18.12±7.89	14.28±2.58	22.41±1.17
PICO	EBM-NLP‡	77.66±0.10	**78.47±0.10**	78.14±0.74	73.78±1.65	77.31±0.01	73.91±0.74
REL	ChemProt†	**83.83±0.55**	82.64±0.32	82.83±1.76	75.04±5.75	79.52±0.37	70.99±1.38
	SciERC‡	**81.23±1.61**	72.10±14.1	77.10±5.74	56.58±17.6	68.23±2.85	73.14±1.00
CLS	CitationIntent‡	**64.46±0.46**	59.08±3.52	47.33±15.3	29.58±6.78	44.37±3.50	50.30±6.81
	MAG‡	73.01±0.09	73.59±0.05	**74.61±0.29**	74.12±0.65	74.23±0.09	72.27±0.05
	SciCite‡	83.96±0.22	80.75±6.24	**85.49±0.63**	83.91±0.10	84.45±0.15	84.22±0.18
DEP	GENIA (LAS)¶	**53.33±0.72**	53.23±1.88	36.31±8.55	24.15±9.25	34.15±2.80	30.08±0.97
	GENIA (UAS)§	57.88±0.75	**58.57±2.22**	41.68±8.60	26.62±4.52	40.44±2.93	34.55±1.33
	Average	**74.91±0.14**	73.55±1.77	71.19±3.97	59.47±4.94	65.50±0.99	65.51±0.55

Table 3: Performance comparison (%) of our models (SciLaD-M) and existing language models on five tasks and eleven datasets. Our SciLaD-M (custom) and SciLaD-M (RoBERTa) models are equipped with a tokenizer trained on our dataset and the tokenizer from existing RoBERTa-base model, respectively. The scores are the F1 scores from each task, averaged over three runs. Notation: † micro-averaged F1 score; ‡ macro-averaged F1 score; ¶ labeled attachment score; § unlabeled attachment score.

5. Related Work

Hu et al. (2025) published a comprehensive survey on current LLMs and datasets for the scientific domain, covering comparable datasets and LLMs, which we discuss in this section. Various transformer-based LLMs have been trained on curated scientific corpora, yielding exceptional results on their domain-specific benchmarks (Beltagy et al., 2019; Phan et al., 2021; Taylor et al., 2022). These models are trained using scientific data from online databases such as arXiv[7], PMC[8], ACL Anthology[9], or Semantic Scholar[10]. However, to the best of our knowledge, the articles from Unpaywall have not been used for LLM pre-training to date. For instance, the encoder-based model SciBERT has been trained on a set of papers from Semantic Scholar (Ammar et al., 2018) and evaluated on downstream tasks like NER, PICO, CLS, REL, and DEP (Beltagy et al., 2019). The sequence-to-sequence model SciFive (Phan et al., 2021) has been trained on a collection of PMC Abstracts and full texts[11]. Larger and more recent models, e. g., GALACTICA (Taylor et al., 2022), use a collection of multiple corpora from arXiv, PMC, ACL Anthology, and Semantic Scholar. They claim to use a corpus with approximately 88 billion tokens. Note that *none* of the data splits used for training these scientific (L)LMs are publicly available, which makes it impossible to analyze the details of the data coverage.

Based on the previously referenced databases for scientific research, there exist four recent and extensive corpora. The largest publicly available corpus of scientific text is S2ORC (Lo et al., 2020), spanning 81.1 million open-access papers from multiple academic fields and curated from various digital archives. S2ORC is constructed from the PDF files of the Semantic Scholar literature corpus, processed using Grobid, and assembled using the metadata provided by Grobid, for instance, by checking the DOI against other open-access databases, such as Unpaywall (Chawla, 2017). UnarXive (Saier et al., 2023) is a dataset based on publications uploaded to arXiv, reaching over 1.9 million documents. UnarXive also covers multiple scientific disciplines and is not limited to a single scientific field. The ACL Anthology Network (AAN) (Radev et al., 2009) is a manually curated, networked database of documents built upon the ACL Anthology. ANN provides a bibliometric-enhanced corpus covering 24.6k papers from computational linguistics and NLP. The fourth corpus is the PMC OA Subset, a dataset comprising journal articles and preprints from PMC. It is important to note that not all articles within PMC are eligible for text mining or other forms of reuse, as a significant proportion remains protected under copyright restrictions. Our contribution extends the current set of public corpora with new datapoints that have not been used for LLM pretraining so far.

[7] https://arxiv.org

[8] https://pmc.ncbi.nlm.nih.gov/tools/openftlist/

[9] https://aclanthology.org

[10] https://www.semanticscholar.org

[11] https://pubmed.ncbi.nlm.nih.gov

6. Conclusion

We present SciLaD, a new dataset of scientific publications, constructed entirely using open-source frameworks and publicly accessible data sources. The dataset, together with a transparent and reproducible pipeline, demonstrates that large-scale scientific corpora can be developed without reliance on proprietary infrastructure or restricted resources. By harvesting open-access texts from Unpaywall, arXiv, and PLOS, our processing and normalization workflow establishes a scalable, ethical blueprint for data creation in natural scientific language processing (NSLP).

The way the Unpaywall OA list is exploited, and the modular approach in the pipeline, allow SciLaD to be incrementally maintained and extended with new scientific articles without rerunning the entire harvesting and preprocessing pipeline from scratch.

Beyond the dataset, we introduce an extensible evaluation pipeline that enables fair and consistent benchmarking across diverse model architectures, including encoder-only, decoder-only, and sequence-to-sequence models. By making both the dataset and the evaluation framework openly available, we aim to support the broader community in advancing domain-specific NLP research. This contribution underscores the importance of transparency, accessibility, and reproducibility as guiding principles for future scientific language technologies.

Limitations

This work has several limitations that should be addressed in future releases of the dataset.

Open Source Data Split While the dataset presented in this work is fully derived from open-access resources, the current release focuses primarily on the textual content and associated metadata. In future iterations, we aim to provide a clearly defined CC-BY data split that ensures unrestricted use and redistribution across academic and industrial settings. This release will also include additional per-sample information, such as corresponding PDF file versions of abstracts, enabling richer downstream tasks such as layout analysis and document structure modeling. Moreover, due to language distribution imbalance, the plain text clean split in this study is limited to English, and expanding it to include other languages would enhance its applicability across diverse linguistic contexts.

Data Curation Although our dataset construction pipeline captures a wide variety of scientific publications, there remains room for improvement in terms of fine-grained curation. In particular, future work will extend entity and relation extraction, as well as structured table and formula parsing, to support more complex information retrieval and reasoning tasks. These improvements are crucial for developing high-quality datasets that can facilitate scientific table-to-text generation and enhance the training of specialized scientific LLMs.

LLM Training Our current experiments focus on discriminative tasks such as classification, relation extraction, and named entity recognition. However, the dataset also holds potential for generative model development. As part of our ongoing work, we plan to train sequence-to-sequence and decoder-only language models—such as compact architectures in the style of *SmolLM2* (Allal et al., 2025) to explore the dataset's applicability for summarization, text generation, and scientific question answering. This direction will enable a broader evaluation of how open, domain-specific corpora can support scalable and transparent LLM training.

Citation graph exploitation The data are represented in TEI XML, which preserves the document structure and citation graphs. Those TEI files contain valuable information that can be used for pretraining LLMs or for fine-tuning of downstream tasks, for instance. However, in our project, we pre-trained the LLMs only on plain text. Moreover, we evaluated our model on 12 downstream biomedical tasks. Those benchmarks evaluate on major NLP tasks in the scientific domain, such as NER, REL, and CLS. Although we believe the results demonstrate improvements in scientific language modeling with our dataset, the focus could be on more recent tasks, such as long-context reasoning, full-document summarization, and sophisticated citation graph modeling. Training a language model on TEI XML would most likely also improve performance on these tasks. Due to computational constraints, we aim to use the data for a future project to train a more sophisticated LLM that will then be evaluated on more recent downstream tasks. In future work, we envision a targeted exploitation of the citation graph as well as keyword mapping, which would help to assess and improve the quality of the data in the following releases.

Data and code availability

The data presented in this article consists of 35.3 million documents structured in TEI XML, which is limited to research activities (https://huggingface.co/datasets/scilons/SciLaD-all-xml-v1). The full text-based version of the dataset is available at https://huggingface.co/datasets/

scilons/SciLaD-all-text-v1 The English deduplicated version is available as parquet at https://huggingface.co/datasets/scilons/SciLaD-en-dedup-v1. A split consisting of only Creative Commons licenses that allow redistribution (see above) will be shared separately without limitations in a future release.

The code is available on GitHub. The harvesting project is available at https://github.com/kermitt2/article_dataset_builder. The evaluation pipeline is hosted at https://github.com/scilons/scilons-eval, the processing pipeline is available at https://github.com/scilons/harvesting, the training code and scripts for SciLaD-M (custom) and SciLaD-M (RoBERTa) are available at https://github.com/scilons/roberta-pretrain. The fork of LaTeXML used in this work is available at https://github.com/kermitt2/LaTeXML/. The rest of the tools used in this work, such as Grobid (GROBID contributors, 2008 - 2025), Pub2TEI (Patrice Lopez, 2014 - 2025) are available in the cited repository.

The pre-trained models are available on Huggingface, the SciLaD-M (custom) final version is at https://huggingface.co/scilons/SciLaD-M-custom and SciLaD-M (RoBERTa) is at https://huggingface.co/scilons/SciLaD-M-roberta.

All checkpoints of SciLaD-M (custom) are available at https://huggingface.co/collections/scilons/scilad-m-custom, the custom tokenizer is available at (https://huggingface.co/scilons/sciroberta-tokenizer-v3. All checkpoints of SciLaD-M (RoBERTa) are available at https://huggingface.co/collections/scilons/scilad-m-roberta.

7. Acknowledgments

The data collection was provided by James Howison and Patrice Lopez in the scope of the SoftCite project[12], aiming to create a large dataset of ML-identified mentions of software (Howison et al., 2025). The harvesting pipeline was run through the JetStream 2 program at Indiana University[13] ran through allocation CIS220172 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The authors also acknowledge the Texas Advanced Computing Center (TACC)[14] at The University of Texas at Austin for providing computational resources that have contributed to the creation and processing of this research dataset.

This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[15] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The consortium is funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). Further support was provided from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101189745 (HIVEMIND).

Finally, the authors acknowledge the University of Mannheim and the Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI) for providing the necessary resources required for training and evaluating the models.

[12] https://github.com/softcite

[13] https://jetstream-cloud.org/

[14] http://www.tacc.utexas.edu

[15] https://www.nfdi4datascience.de

Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Dalmeet Singh Chawla. 2017. Unpaywall finds free versions of paywalled papers. *Nature*.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78.

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

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Holly Else. 2018. How unpaywall is transforming open science. *Nature*, 560(7718):290–292.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Deyan Ginev, Bruce R. Miller, and Silviu Oprea. 2014. E-books and graphics with latexml.

GROBID contributors. 2008 - 2025. Grobid (generation of bibliographic data). https://github.com/kermitt2/grobid. Swh:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

James Howison, Patrice Lopez, Ram Karthik, and Will Beason. 2025. Softcite extractions from the open access literature.

Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, Ying Chen, Chaoyang Zhang, Cheng Tan, Jie Ying, Guocheng Wu, Shujian Gao, Pengcheng Chen, Jiashi Lin, Haitao Wu, Lulu Chen, Fengxiang Wang, Yuanyuan Zhang, Xiangyu Zhao, Feilong Tang, Encheng Su, Junzhi Ning, Xinyao Liu, Ye Du, Changkai Ji, Pengfei Jiang, Cheng Tang, Ziyan Huang, Jiyao Liu, Jiaqi Wei, Yuejin Yang, Xiang Zhang, Guangshuai Wang, Yue Yang, Huihui Xu, Ziyang Chen, Yizhou Wang, Chen Tang, Jianyu Wu, Yuchen Ren, Siyuan Yan, Zhonghua Wang, Zhongxing Xu, Shiyan Su, Shangquan Sun, Runkai Zhao, Zhisheng Zhang, Dingkang Yang, Jinjie Wei, Jiaqi Wang, Jiahao Xu, Jiangtao Yan, Wenhao Tang, Hongze Zhu, Yu Liu, Fudi Wang, Yiqing Shen, Yuanfeng Ji, Yanzhou Su, Tong Xie, Hongming Shan, Chun-Mei Feng, Zhi Hou, Diping Song, Lihao Liu, Yanyan Huang, Lequan Yu, Bin Fu, Shujun Wang, Xiaomeng Li, Xiaowei Hu, Yun Gu, Ben Fei, Benyou Wang, Yuewen Cao, Minjie Shen, Jie Xu, Haodong

Duan, Fang Yan, Hongxia Hao, Jielan Li, Jiajun Du, Yanbo Wang, Imran Razzak, Zhongying Deng, Chi Zhang, Lijun Wu, Conghui He, Zhaohui Lu, Jinhai Huang, Wenqi Shao, Yihao Liu, Siqi Luo, Yi Xin, Xiaohong Liu, Fenghua Ling, Yuqiang Li, Aoran Wang, Siqi Sun, Qihao Zheng, Nanqing Dong, Tianfan Fu, Dongzhan Zhou, Yan Lu, Wenlong Zhang, Jin Ye, Jianfei Cai, Yirong Chen, Wanli Ouyang, Yu Qiao, Zongyuan Ge, Shixiang Tang, Junjun He, Chunfeng Song, Lei Bai, and Bowen Zhou. 2025. A survey of scientific large language models: From data foundations to agent frontiers.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics (Oxford, England)*, 19 Suppl 1:i180–2.

Petr Knoth, Drahomira Herrmannova, Matteo Cancellieri, Lucas Anastasiou, Nancy Pontika, Samuel Pearce, Bikash Gyawali, and David Pride. 2023. Core: A global aggregation service for open access papers. *Scientific Data*, 10(1):366.

Jens Vindahl Kringelum, Sonny Kim Kjærulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database: The Journal of Biological Databases and Curation*, 2016.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197.

Laurent Romary Patrice Lopez. 2014 - 2025. Pub2tei. https://github.com/kermitt2/pub2tei. Swh:1:dir:291d25e23860f47dbf49ab6c2a18c1a2731fecea.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The fineweb datasets: Decanting the web for the finest text data at scale.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024b. Datatrove: large scale data processing.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature.

Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. 2018. The state of oa: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, 6:e4375.

Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

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

Tarek Saier, Johan Krause, and Michael Färber. 2023. unarxive 2022: All arxiv publications preprocessed for nlp, including structured full-text and citation network. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, page 66–70. IEEE.

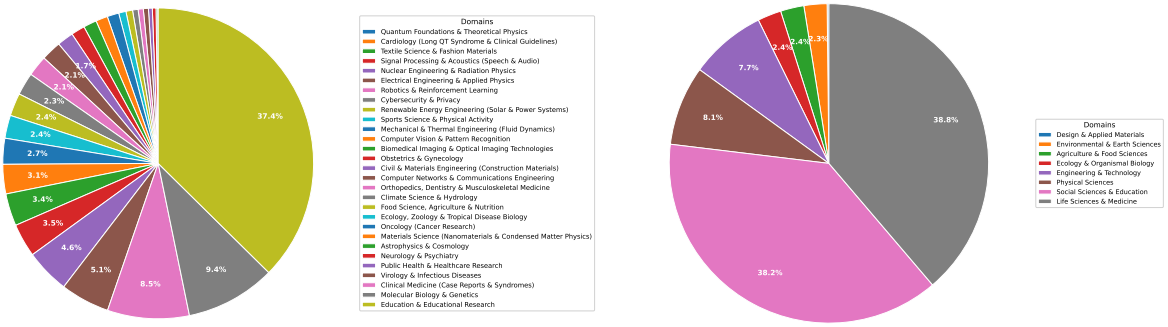
Elizabeth Seiver, M Pacer, and Sebastian Bassi. 2018. Text and data mining scientific articles with allofplos. In *Proceedings of the 17th Python in Science Conference*, pages 61–64.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. 2017. Text mining of 15 million full-text scientific articles.



(a) Fine-grained domain distribution

(b) High-level domain distribution

Figure 3: Distribution of domains in the full dataset.

A. Domain Distribution

To showcase the available scientific domains in the dataset, we apply a two-stage topic modeling approach. First, a representative sample of one million paper titles is drawn from the dataset using reservoir sampling, ensuring a uniform sample of the corpus. Each title was then encoded using SentenceTransformers[16], and all embeddings were given to BERTopic (Grootendorst, 2022) to produce 30 fine-grained topic clusters. We then labeled the resulting topics using GPT5 (Singh et al., 2025) based on the keywords and representative document produced by BERTopic. The topics were manually reviewed and clustered further into 8 high-level scientific domains. Once the topic model was trained, we applied a second straming pass across the full corpus, assigning each title to the closest cluster using the same embedding method. Figure 3a shows the fine-grained topic distribution and Figure 3b the high-level domains in SciLaD.

[16]Specifically: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2