MultiLegalPile: A 689GB Multilingual Legal Corpus

Joel Niklaus [1 2 3] **Veton Matoshi** [2] **Matthias Stürmer** [1 2] **Ilias Chalkidis** [4] **Daniel E. Ho** [3]

Abstract

Large, high-quality datasets are crucial for training Large Language Models (LLMs). However, so far, there are few datasets available for specialized critical domains such as law and the available ones are often only for the English language. We curate and release MULTILEGALPILE, a 689GB corpus in 24 languages from 17 jurisdictions. The MULTILEGALPILE corpus, which includes diverse legal data sources with varying licenses, allows for pretraining NLP models under fair use, with more permissive licenses for the Eurlex Resources and Legal mC4 subsets. We pretrain two RoBERTa models and one Longformer multilingually, and 24 monolingual models on each of the language-specific subsets and evaluate them on LEXTREME. Additionally, we evaluate the English and multilingual models on LexGLUE. Our multilingual models set a new SotA on LEXTREME and our English models on LexGLUE. We release the dataset, the trained models, and all of the code under the most open possible licenses.

1. Introduction

Recent years have seen LLMs achieving remarkable progress, as demonstrated by their performance on various benchmarks such as SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), and several human Exams (OpenAI, 2023), including U.S. bar exams for admission to practicing the law (Katz et al., 2023). These models are typically trained on increasingly large corpora, such as the Pile (Gao et al., 2020a), C4 (Raffel et al., 2020b), and mC4 (Xue et al., 2021). However, it is important to note that public corpora available for training these models are predominantly in English, and often constitute web text with unclear licensing. This even led to lawsuits against LLM producers, highlighting this critical issue. Furthermore, there is a scarcity of large-scale, domain-specific pretraining corpora, which constitutes a significant gap in the current body of resources available for the training of LLMs. Similarly, LLMs are predominantly English, especially considering domain-specific models, e.g., ones specialized in biomedical, legal, or financial texts.

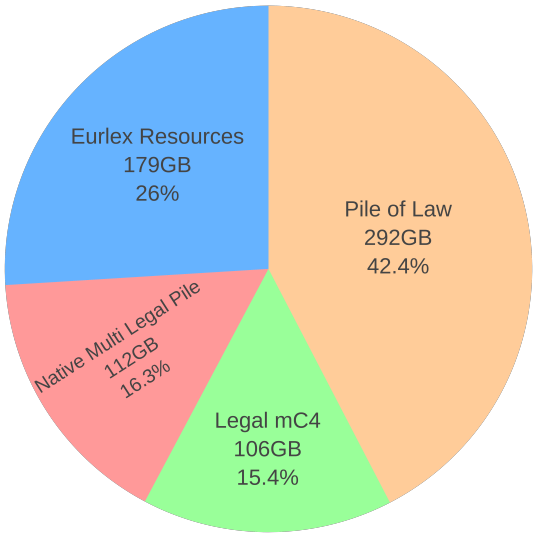

Figure 1. MULTILEGALPILE Source Distribution

Legal texts, often produced by public instruments (e.g., state governments, international organizations), are typically available under public licenses, offering a rich resource for domain-specific pretraining. Given this context, we curate a humongous, openly available, corpus of multilingual law text spanning across numerous jurisdictions (legal systems), predominantly under permissive licenses.

Further on, we continue pretraining XLM-R models (Conneau & Lample, 2019) on our corpus and evaluated these models on the recently introduced LEXTREME (Niklaus et al., 2023) and LexGLUE (Chalkidis et al., 2021e) benchmarks. Given the often extensive nature of legal text, we also pretrained a Longformer model (Beltagy et al., 2020) for comparison with hierarchical models (Chalkidis et al., 2019b; Niklaus et al., 2021; 2022).

Our multilingual models set new SotA on LEXTREME overall. Our legal Longformer outperforms all other models in four LEXTREME datasets and reaches the highest dataset aggregate score. Our monolingual models outperform their

[1]Institute of Computer Science, University of Bern, Bern, Switzerland [2]Bern University of Applied Sciences [3]Stanford University [4]University of Copenhagen. Correspondence to: Joel Niklaus <joel.niklaus@unibe.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

base model XLM-R in 21 out of 24 languages, even reaching language specific SotA in five. On LexGLUE our English models reach SotA in five out of seven tasks with the large model achieving the highest aggregate score.

In the spirit of open science, we provide the dataset under a CC BY-NC-SA 4.0 license, with some subsets licensed more permissively. Dataset creation scripts, models, and pretraining code are public under Apache 2.0 licenses. This open-source approach encourages further research and advancements in the field of legal text analysis and understanding using large language models.

Contributions

The contributions of this paper are three-fold:

1. We curate and release a large scale multilingual legal text corpus, dubbed MULTILEGALPILE,[1] covering 24 languages and 17 legal systems (jurisdictions).
2. We release 2 multilingual and 24 monolingual new legal-oriented PLMs, dubbed LEGALXLMS, warm-started from the XLM-R (Conneau & Lample, 2019) models, and further pretrained on the MULTILEGALPILE. Additionally, we pretrain a Longformer (Beltagy et al., 2020) based on our multilingual base-size model on context lengths of up to 4096 tokens.
3. We benchmark the newly released models on the LEXTREME and LexGLUE benchmarks, achieving new SotA for base and large size models and increasing performance drastically in Greek legal code. Our Longformer model reaches SotA in four tasks and the highest dataset aggregate score. Our monolingual models set language specific SotA in five languages.

2. Related Work

2.1. General Pretraining Corpora

The use of pretrained Language Models (PLMs) has become increasingly popular in NLP tasks, particularly with the advent of models such as BERT (Devlin et al., 2019) that can be finetuned for specific applications. One key factor in the success of pretraining is the availability of large and diverse text corpora, which can help the model learn the nuances of natural language. In the following, we discuss large-scale general-purpose text corpora used for pretraining.

Wikipedia is a commonly used multilingual dataset for pretraining language models, and has been used to pretrain BERT (Devlin et al., 2019), MegatronBERT (Shoeybi et al., 2020), T5 (Raffel et al., 2020a), and GPT-3 (Brown et al., 2020b), among others.

Based on Wikipedia, Merity et al. (2016) created WikiText by selecting articles fitting the Good or Featured article criteria. The dataset contains 103M words and has two versions: WikiText2 and the larger WikiText103. It has been used to pretrain models like MegatronBERT (Shoeybi et al., 2020) and GPT-2 (Radford et al., 2019).

The BookCorpus (Zhu et al., 2015), also known as the Toronto Books Corpus, is an English dataset used for pretraining BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020a). It consists of almost 1B words from over 11K books collected from the web.

The Common Crawl corpus is a publicly available multilingual dataset of scraped web pages, regularly updated with new "snapshots". It has been used to pretrain GPT-3 (Brown et al., 2020b) as well as XLM-R (Conneau et al., 2020a). One significant drawback of Common Crawl is the presence of uncleaned data, which includes a considerable amount of "gibberish or boiler-plate text like menus, error messages, or duplicate text" (Raffel et al., 2020a). As a result, utilizing the Common Crawl dataset necessitates additional post-filtering and cleaning procedures. To address this issue, Raffel et al. (Raffel et al., 2020a) performed several cleaning steps on the April 2019 snapshot of Common Crawl, resulting in the creation of the Colossal Clean Crawled Corpus (C4), comprising 750 GB of English-language text. It was used for pretraining models such as T5 (Raffel et al., 2020a) and Switch Transformer (Fedus et al., 2022).

OpenWebText (Gokaslan & Cohen, 2019) openly replicates OpenAI's closed English WebText dataset (Radford et al., 2019), used to pretrain GPT-2 (Radford et al., 2019). WebText comprises over 8M documents with a combined text size of 40 GB. To ensure data uniqueness, any documents sourced from Wikipedia were excluded from WebText, as they are commonly utilized in other datasets. OpenWebText, on the other hand, consists of 38 GB of text data from 8M documents and was used for pretraining RoBERTa (Liu et al., 2019) and MegatronBERT (Shoeybi et al., 2020).

News articles are also a common source for pretraining corpora. The RealNews dataset (Zellers et al., 2019) is a large corpus extracted from Common Crawl, containing news articles from December 2016 to March 2019 (training) and April 2019 (evaluation), totaling 120 GB. It was used for pretraining MegatronBERT (Shoeybi et al., 2020). For pretraining RoBERTa, Liu et al. (2019) used an English subset of RealNews, comprising 63M English news articles crawled from September 2016 to February 2019.

The rise of LLMs brought about the creation of ever larger training datasets. The Pile (Gao et al., 2020b) combines 22 distinct, well-curated datasets, such as Wikipedia (English), OpenWebText2 (Gokaslan & Cohen, 2019), OpenSubtitles (Tiedemann, 2016) etc., encompassing 825 GB of data. Besides general-purpose textual datasets, it also

[1] https://huggingface.co/datasets/joelito/Multi_Legal_Pile

contains domain-specific datasets, such as ArXiv (Science), FreeLaw (Legal), PubMed Abstracts (Biomedicine), and GitHub data (to improve code-related task performance (Gao et al., 2020b)). GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020b) were evaluated on this dataset.

In their work, Touvron et al. (2023) compiled a substantial dataset from various publicly available sources, including CommonCrawl, C4, Github, Wikipedia, etc., totaling 1.4T tokens. They trained the 13B-parameter LLaMA model using this dataset, surpassing the performance of the 175B-parameter GPT-3 on most benchmark tasks. However, the dataset itself is not publicly available. To address this, a collaborative effort resulted in the creation of the RedPajama-Data-1T dataset, replicating LLaMA's dataset with a similar size of 1.2T tokens.

Some of the afore-mentioned datasets, such as Common Crawl, are used to pretrain multilingual versions of BERT, DistilBERT, RoBERTa etc. These models were pretrained on datasets that cover approximately 100 languages, thereby neglecting low-resource languages. ImaniGooghari et al. (2023) addressed this by compiling Glot500, a 700 GB dataset covering 500 diverse languages, with a focus on low-resource ones. The Glot500-m model, pretrained on this dataset, outperformed the XLM-RoBERTa base model on six out of seven tasks.

2.2. Domain Specific Corpora

While pretraining on general-purpose text like Wikipedia and news articles shows promise, evidence suggests that pretraining on domain-specific text can enhance language model performance on related tasks (Beltagy et al., 2019; Gu et al., 2021; Chalkidis et al., 2020b; Niklaus & Giofré, 2022). Domain-specific text corpora include texts specific to fields like medicine, law, or science.

Several studies have examined pretraining on scientific text corpora. Beltagy et al. (2019) pretrained SciBERT, a BERT-based model, on a random subset of 1.14M papers sourced from Semantic Scholar. This collection comprises 18% of computer science papers and 82% of papers from the broader biomedical field. Similarly, PubMed and PubMedCentral are common sources for biomedical datasets. Gu et al. (2021) trained PubMedBERT using PubMed abstracts and PubMedCentral articles; BioBERT (Lee et al., 2020) was pretrained similarly. Johnson et al. (2016) compiled the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, a large single-center database of critical care patients. Huang et al. (2019) used over 2 million de-identified clinical notes from this dataset to pretrain ClinicalBERT. These models outperformed general-purpose models on biomedical NLP tasks.

In the legal domain, similar strategies are observed. Chalkidis et al. (2020a) collected 12 GB of diverse English legal texts, including legislation, court cases, and contracts. They pretrained LegalBERT on this dataset, showing state-of-the-art performance, especially in tasks requiring domain knowledge. Another study by Zheng et al. (2021) used the entire English Harvard Law case corpus (1965-2021) comprising 37 GB of text to pretrain CaseLaw-BERT.

Recently, Chalkidis* et al. (2023) released LexFiles, an English legal corpus with 11 sub-corpora covering legislation and case law from six English-speaking legal systems (EU, Council of Europe, Canada, US, UK, India). The corpus contains approx. 6M documents or approx. 19B tokens. They trained two new legal English PLMs, showing improved performance in legal probing and classification tasks.

Efforts to pretrain legal language models also exist for Italian (Licari & Comandè, 2022), Romanian (Masala et al., 2021), and Spanish (Gutiérrez-Fandiño et al., 2021). However, English dominates, underscoring the importance of compiling multilingual legal corpora.

Model	Domain	Languages	Size in # Words
SciBERT (Beltagy et al., 2019)	scientific	English	2.38B (3.17B tokens)
Galactica (Taylor et al., 2022)	scientific	English	79.5B (106B tokens)
BioBERT (Lee et al., 2019)	biomedical	English	18B
LegalBERT (Chalkidis et al., 2020b)	legal	English	1.44B (11.5GB)
CaselawBERT (Zheng et al., 2021)	legal	English	4.63B (37GB)
LegalXLMs (ours)	legal	24 EU langs	87B (689GB)

Table 1. Previous domain specific pretraining corpora. For some corpora only GB or tokens were available. We converted 8 GB into 1B words and 1 token to 0.75 words.

Table 1 compares previous domain-specific corpora, all in English. In terms of size, none reach the MULTILEGALPILE proposed here.

3. MULTILEGALPILE

3.1. Construction

We transformed all datasets into xz compressed JSON Lines (JSONL) format. The combination of XZ compression and JSONL is ideal for streaming large datasets due to reduced file size and efficient decompression and reading.

Filtering mC4 We employed the vast multilingual web crawl corpus, mC4 (Xue et al., 2021), as our foundation. To effectively filter this corpus for legal content, we utilized regular expressions to identify documents with legal references. We found that detecting legal citations, such as references to laws and rulings, served as a reliable indicator of legal-specific documents in the corpus.

In order to ensure the accuracy of our filtering, we engaged legal experts to aid in identifying citations to laws and rulings across different jurisdictions and languages. We manually reviewed the precision of the retrieved documents for five languages, namely German, English, Spanish, French,

Iteration	German	English	Spanish	French	Italian
1st	100%	20%	100%	65%	80%
2nd	100%	85%	100%	100%	95%

Table 2. Precision of investigated languages in legal mC4 (n=20)

and Italian, as shown in Table 2. The proficiency levels of the evaluators included native German, fluent English and Spanish, intermediate French, and basic Italian.

Subsequent to the initial review, we performed a second round of precision evaluation, during which we refined our regex expressions based on our findings from the first iteration. This iterative process not only enhanced the precision of the legal content detection, but also resulted in a reduction of the corpus size from 133GB to 106GB. Although the overall volume of data was reduced, this process significantly improved the quality and specificity of the corpus by focusing on legal content with a higher degree of precision.

A major reason for utilizing regexes instead of a Machine Learning (ML) based classifier was speed. Already when utilizing regexes, filtering through such a huge corpus like mC4 (27TB in total, of which 10.4TB are in English) took several days. An ML model based on Bag-of-Words, Word vectors or even contextualized embeddings would a) need an annotated dataset and b) likely be much slower.

Figure 2. MULTILEGALPILE Text Type Distribution

Compiling Native MULTILEGALPILE To compile the corpus, we scraped several sources containing legal language materials. Our search was conducted in a loose manner, meaning that when we found a suitable source with legal text data, we included it in our corpus. It is important to note that we do not claim completeness, as we were unable to perform quality analysis for all available languages. For a detailed overview of sources used for the Native MULTILEGALPILE corpus, please refer to Table 9.

The majority of sources provided a link to download the data directly. In cases where data was formatted differently, we converted it into a unified format, such as jsonl. The post-processing steps involved performing various tasks depending on the initial data format. For example, in the case of CASS, we extracted the textual data from XML tags.

Curating Eurlex Resources To curate the Eurlex resources, we utilized the eurlex R package to generate SPARQL queries and download the data. Subsequently, we converted the data into a format more amenable to handling large datasets using Python.

Integrating Pile of Law Henderson et al. (2022) released a large corpus of diverse legal text in English mainly originating from the US. We integrated the latest version with additional data (from January 8, 2023) into our corpus.

3.2. Description

MULTILEGALPILE consists of four large subsets: a) Native Multi Legal Pile (112 GB), b) Eurlex Resources[2] (179 GB), c) Legal MC4[3] (106 GB) and d) Pile of Law (Henderson et al., 2022) (292 GB).

Figure 3 details the distribution of languages. Note that due to the integration of the Pile of Law, English is by far the most dominant language, representing over half of the words. In Figure 2 we show the distribution across text types. Caselaw makes up over half of the corpus, due to the good public access to court rulings especially in common law countries. Note, that even in civil law countries — where legislation is much more important – caselaw is usually more plentiful than legislation (as can be seen in the Swiss case in Table 9). It is hard to find publicly available contracts, leading to the relatively low percentage of the total corpus ($< 10\%$), even though they could potentially make up most of the legal texts in existence (from the private sector). Note that most of the contracts in our corpus are from the US or international treaties with the EU. Table 9 in Appendix C provides additional of the MULTILEGALPILE, including sources and licenses.

3.3. Licenses and Usage of MULTILEGALPILE

The Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license applied for the released MULTILEGALPILE corpus depends on the upstream licenses of the data subsets described above.

First, our *Native Multi Legal Pile* consists of data sources with different licenses. They range from restrictive licenses such as CC BY-NC-SA 4.0 up to the most liberal Creative Commons Zero (CC0) license, which, in essence, releases

[2] https://huggingface.co/datasets/joelito/eurlex_resources
[3] https://huggingface.co/datasets/joelito/legal-mc4

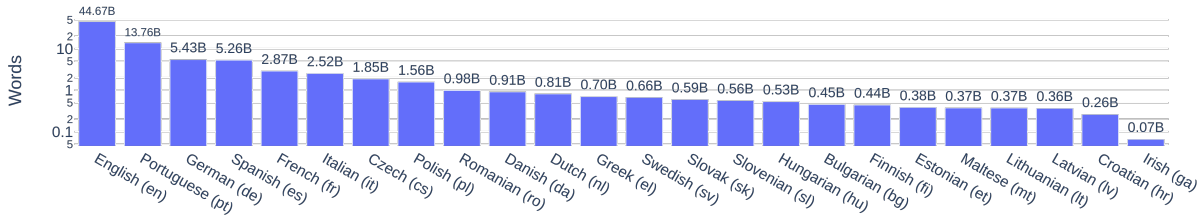

Figure 3. MULTILEGALPILE Language Distribution (Note the log-scaled y-axis)

the data into the public domain. Many sources, however, do not explicitly state the license used for the available data. We assume that such data sources allow pretraining usage, since the creators are usually public agencies such as courts and administrations. Such legislation and caselaw is usually not protected by copyright law. Table 9 provides an overview of the license or copyright situation for each of the 29 sources in the Native Multi Legal Pile.

Second, the *Eurlex Resources* is licensed under CC BY 4.0 by the European Union.[4] Thus, including this corpus does not pose legal issues for pretraining.

Third, the *Legal mC4* corpus was created by filtering multilingual C4 (Xue et al., 2021) for legal content as described above. As mC4 is licensed under ODC-BY, we also release the filtered Legal mC4 corpus under the same license.

Finally, the *Pile of Law* (Henderson* et al., 2022) is published under CC BY-NC-SA 4.0 and the dataset is not altered, therefore the license remains the same.

Usage of the MULTILEGALPILE corpus is presumably possible for pretraining of NLP models. In general, we assume that the fair use doctrine allows employing the data for legal NLP models because the results are rather transformative (Henderson et al., 2023). Nevertheless, copyright issues in generative AI remain an unresolved problem for the moment. Several court cases are currently pending, such as Getty Images suing Stability AI for intellectual property infringement (Sag, 2023).

4. Pretraining Legal Models

As part of this study, we release 2 new multi-lingual legal-oriented PLMs, dubbed Legal-XLM-Rs, trained on the newly introduced MULTILEGALPILE corpus (Section 3). For the newly released Legal-XLM-Rs we followed a series of best-practices in language model development literature:

(a) We warm-start (initialize) our models from the original XLM-R checkpoints (base or large) of Conneau & Lample (2019). Model recycling is a standard process followed by many (Wei et al., 2021; Ouyang et al., 2022) to benefit from starting from an available "well-trained" PLM, rather from scratch (random). XLM-R was trained on 2.5TB of cleaned CommonCrawl data in 100 languages.

(b) We train a new tokenizer of 128K BPEs on the training subsets of MULTILEGALPILE to better cover legal language across all available legal systems and languages. However, we reuse the original XLM-R embeddings for all lexically overlapping tokens (Pfeiffer et al., 2021), i.e., we warm-start word embeddings for tokens that already exist in the original XLM-R vocabulary, and use random ones for the rest.

(c) We continue pretraining our models on the diverse MULTILEGALPILE corpus with batches of 512 samples for an additional 1M/500K steps for the base/large model. We do initial warm-up steps for the first 5% of the total training steps with a linearly increasing learning rate up to $1e-4$, and then follow a cosine decay scheduling, following recent trends. For half of the warm-up phase (2.5%), the Transformer encoder is frozen, and only the embeddings, shared between input and output (MLM), are updated. We also use an increased 20/30% masking rate for base/large models respectively, where also 100% of the predictions are based on masked tokens, compared to Devlin et al. (2019)[5], based on the findings of Wettig et al. (2023).

(d) For both training the tokenizer and our legal models, we use a sentence sampler with exponential smoothing of the sub-corpora sampling rate following Conneau & Lample (2019) and Raffel et al. (2020b), since there is a disparate proportion of tokens across sub-corpora and languages (Figures 1 and 3) and we aim to preserve per-corpus and language capacity, i.e., avoid overfitting to the majority (approx. 50% of the total number of tokens) US-origin English texts.

(e) We consider mixed cased models, i.e., both upper- and lowercase letters covered, similar to all recently developed large PLMs (Conneau & Lample, 2019; Raffel et al., 2020b; Brown et al., 2020a).

To better account for long contexts often found in legal documents, we continue training the base-size multilingual model on long contexts (4096 tokens) with windowed attention (128 tokens window size) (Beltagy et al., 2020) for 50K steps, dubbing it Legal-XLM-LF-base. We use the standard 15% masking probability and increase the learning rate to

[4] EUR-Lex Legal notice

[5] Devlin et al. – and many other follow-up work – used a 15% masking ratio, and a recipe of 80/10/10% of predictions made across masked/randomly-replaced/original tokens.

Model	Source	Params	Vocab	Specs	Corpus	# Langs
MiniLM	Wang et al. (2020)	118M	250K	1M steps / BS 256	2.5TB CC100	100
DistilBERT	Sanh et al. (2020)	135M	120K	BS up to 4000	Wikipedia	104
mDeBERTa-v3	He et al. (2021b;a)	278M	128K	500K steps / BS 8192	2.5TB CC100	100
XLM-R base	Conneau et al. (2020b)	278M	250K	1.5M steps / BS 8192	2.5TB CC100	100
XLM-R large	Conneau et al. (2020b)	560M	250K	1.5M steps / BS 8192	2.5TB CC100	100
Legal-XLM-R-base	ours	184M	128K	1M steps / BS 512	689GB MLP	24
Legal-XLM-R-large	ours	435M	128K	500K steps / BS 512	689GB MLP	24
Legal-XLM-LF-base	ours	208M	128K	50K steps / BS 512	689GB MLP	24
Legal-mono-R-base	ours	111M	32K	200K steps / BS 512	689GB MLP	1
Legal-mono-R-large	ours	337M	32K	500K steps / BS 512	689GB MLP	1

Table 3. Models: All models can process up to 512 tokens, except Legal-XLM-LF-base which can process up to 4096 tokens. BS is short for batch size. MLP is short for MULTILEGALPILE. Params is the total parameter count (including the embedding layer).

$3e-5$ before decaying but otherwise use the same settings as for training the small-context models.

In addition to the multilingual models, we also train 24 monolingual models on each of the language-specific subsets of the corpus. Except for choosing a smaller vocab size of 32K tokens, we use the same settings as for the multilingual models. Due to resource constraints, we only train base-size models and stop training at 200K steps. Due to limited data available in some low-resource languages, these models sometimes do multiple passes over the data. Because of plenty of data and to achieve a better comparison on LexGLUE, we continued training the English model for 1M steps and also trained a large-size model for 500K steps. See Table 7 in appendix A for an overview.

We make all our models publicly available alongside all intermediate checkpoints (every 50K/10K training steps for RoBERTa/Longformer models) on the Hugging Face Hub.[6]

5. Evaluating on LEXTREME and LexGLUE

5.1. Benchmark Description

Below we briefly describe each dataset. We refer the interested reader to the original papers for more details.

LEXTREME (Niklaus et al., 2023) is a multilingual legal benchmark. It includes five single label text classification datasets, three multi label text classification datasets and four Named Entity Recognition (NER) datasets.

Brazilian Court Decisions (BCD) (Lage-Freitas et al., 2022) is from the State Supreme Court of Alagoas (Brazil) and involves predicting case outcomes and judges' unanimity on decisions. **German Argument Mining (GAM)** (Urchs et al., 2021) contains200 German court decisions for classifying sentences according to their argumentative function. **Greek Legal Code (GLC)** (Papaloukas et al., 2021) tackles topic classification of Greek legislation documents. Tasks involve predicting topic categories at volume, chapter, and subject levels. **Swiss Judgment Prediction (SJP)** (Niklaus et al., 2021) focuses on predicting the judgment outcome from 85K cases from the Swiss Federal Supreme Court. **Online Terms of Service (OTS)** (Drawzeski et al., 2021) contains 100 contracts for detecting unfair clauses with the tasks of classifying sentence unfairness levels and identifying clause topics. **COVID19 Emergency Event (C19)** (Tziafas et al., 2021): consists of legal documents from several European countries related to COVID-19 measures where models identify the type of measure described in a sentence. **MultiEURLEX (MEU)** (Chalkidis et al., 2021b) is a corpus of 65K EU laws annotated with EUROVOC taxonomy labels. Task involves identifying labels for each document. **Greek Legal NER (GLN)** (Angelidis et al., 2018) is a dataset for NER in Greek legal documents. **LegalNERo (LNR)** (Pais et al., 2021) tackles NER in Romanian legal documents. **LeNER BR (LNB)** (Luz de Araujo et al., 2018) addresses NER in Brazilian legal documents. **MAPA (MAP)** (Baisa et al., 2016) is a multilingual corpus based on EUR-Lex for NER annotated at a coarse-grained and fine-grained level.

LexGLUE (Chalkidis et al., 2021d) is a legal benchmark covering two single label text classification datasets, four multi label text classification datasets and a multiple choice question answering dataset.

ECtHR Tasks A & B (Chalkidis et al., 2019a; 2021c) contain approx. 11K cases from the European Court of Human Rights (ECtHR) public database. Based on case facts, Task A involves predicting violated articles of the European Convention of Human Rights (ECHR) and Task B involves predicting allegedly violated articles. **SCOTUS** (Spaeth et al.) combines information from US Supreme Court (SCOTUS) opinions with the Supreme Court DataBase (SCDB). The task is to classify court opinions into 14 issue areas. **EUR-LEX** (Chalkidis et al., 2021a) contains 65K EU laws from the EUR-Lex portal, annotated with EuroVoc concepts. The task is to predict EuroVoc labels for a given document. **LEDGAR** (Tuggener et al., 2020) contains approx. 850K contract provisions from the US Securities and Exchange

[6] https://huggingface.co/joelito

Model	BCD	GAM	GLC	SJP	OTS	C19	MEU	GLN	LNR	LNB	MAP	Agg.
MiniLM	53.0	73.3	42.1	67.7	44.1	5.0	29.7	74.0	84.5	93.6	57.8	56.8
DistilBERT	54.5	69.5	62.8	66.8	56.1	25.9	36.4	71.0	85.3	89.6	60.8	61.7
mDeBERTa-v3	60.2	71.3	52.2	69.1	66.5	29.7	37.4	73.3	85.1	94.8	67.2	64.3
XLM-R-base	63.5	72.0	57.4	69.3	67.8	26.4	33.3	**74.6**	**85.8**	94.1	62.0	64.2
XLM-R-large	58.7	73.1	57.4	69.0	**75.0**	29.0	**42.2**	74.1	85.0	**95.3**	68.0	66.1
Legal-XLM-R-base	62.5	72.4	68.9	70.2	70.8	30.7	38.6	73.6	84.1	94.1	**69.2**	66.8
Legal-XLM-R-large	63.3	73.9	59.3	70.1	74.9	**34.6**	39.7	73.1	83.9	94.6	67.3	66.8
Legal-XLM-LF-base	**72.4**	**74.6**	**70.2**	**72.9**	69.8	26.3	33.1	72.1	84.7	93.3	66.2	**66.9**

Table 4. Dataset aggregate scores for multilingual models on LEXTREME. We report macro-F1 and the best scores in bold.

Model	bg	cs	da	de	el	en	es	et	fi	fr	ga	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv	Agg.
MiniLM	52.7	48.6	42.8	54.6	50.3	34.3	40.1	46.3	42.2	39.0	42.8	29.7	29.6	40.5	44.2	40.8	40.8	29.5	22.7	61.6	59.6	44.3	30.0	43.4	40.5
DistilBERT	54.2	48.6	46.0	60.1	58.8	48.0	50.0	48.8	49.6	47.9	51.4	35.9	31.2	50.1	51.9	41.5	44.4	34.6	34.5	63.2	63.8	51.3	36.2	50.1	46.7
mDeBERTa-v3	54.1	51.3	51.7	63.6	57.7	50.7	53.3	50.8	54.6	49.2	54.9	37.4	37.5	55.1	53.9	47.0	52.5	42.1	41.0	65.7	65.3	55.4	37.5	56.1	50.5
XLM-R-base	56.4	48.3	48.3	60.6	57.6	50.1	47.2	46.7	48.6	49.4	50.1	33.6	32.8	53.4	50.0	44.1	43.8	35.2	41.3	66.1	63.7	45.3	33.7	50.0	47.1
XLM-R-large	**59.9**	56.0	**56.3**	65.4	60.8	56.2	**56.6**	56.5	**56.9**	51.4	55.4	42.5	38.1	**58.5**	58.1	49.9	53.9	39.5	**46.4**	**68.6**	**66.8**	**57.9**	42.4	**59.1**	**53.7**
Legal-XLM-R-base	55.6	**58.8**	50.4	63.6	**63.7**	66.8	56.3	**57.0**	52.6	50.1	56.6	38.7	**56.5**	56.1	57.2	49.1	56.0	41.6	43.9	68.2	66.1	55.6	38.6	54.9	53.5
Legal-XLM-R-large	57.8	55.6	50.4	**65.7**	60.7	**69.3**	55.7	54.5	56.6	**53.3**	**57.2**	39.7	39.1	58.1	**60.6**	48.4	57.2	39.4	45.5	67.3	65.5	49.3	39.7	56.4	53.6
Legal-XLM-LF-base	54.4	49.3	48.1	64.0	60.5	52.8	49.2	52.2	48.2	48.5	55.4	33.0	34.7	54.6	54.8	45.2	52.5	40.1	40.6	68.3	64.1	48.4	33.0	51.3	48.9
NativeLegalBERT	-	-	-	-	-	53.1	46.9	-	-	-	-	-	-	45.3	-	-	-	-	-	-	59.0	-	-	-	51.1
NativeBERT	54.8	57.3	51.2	63.0	62.3	52.0	42.6	47.2	52.4	49.4	50.1	-	37.4	47.1	-	-	-	37.0	40.5	66.5	63.1	44.8	-	55.1	50.2
Legal-mono-R-base	55.9	49.5	51.5	61.3	61.3	50.5	52.1	53.5	53.6	51.1	52.2	**44.1**	54.1	51.8	55.5	**50.0**	**59.1**	**54.3**	34.4	67.1	61.5	48.8	**53.4**	58	53.5

Table 5. Language aggregate scores on LEXTREME. We report macro-F1 and best scores in bold. For each language, we also list the best-performing monolingual legal model under *NativeLegalBERT*, the best-performing monolingual non-legal model under *NativeBERT* and our monolingual legal models under *Legal-mono-R-base*. Missing values indicate that no suitable models were found.

Commission (SEC) filings. The task is to classify contract provisions into categories. **UNFAIR-ToS** (Lippi et al., 2019) contains 50 Terms of Service (ToS) from online platforms, annotated with types of unfair contractual terms. The task is to predict unfair types for a given sentence. **CaseHOLD** (Zheng et al., 2021) contains approx. 53K multiple choice questions about holdings of US court cases. The task is to identify the correct holding statement from a selection of five choices.

5.2. Experimental Setup

To ensure comparability, we followed the experimental setups described in the original papers (Niklaus et al., 2023; Chalkidis et al., 2021d) using hierarchical transformers for datasets where the sequence length of most documents exceeds the maximum sequence length of the model (Aletras et al., 2016; Niklaus et al., 2022). The hyperparameters used for running experiments on each dataset are provided in Table 8 in the appendix. To obtain Table 6, we followed Chalkidis et al. (2021d), running five repetitions with different random seeds (1-5) and reporting the test scores based on the seed that yielded the best scores on the development data. For values in Tables 4 and 5, we followed the procedure in Niklaus et al. (2023), taking the mean of the results of 3 random seeds (1-3). We show an overview of the evaluated models in Table 3.

5.3. Evaluation on LEXTREME

We evaluate our models on LEXTREME (Niklaus et al., 2023) and show results across datasets in Table 4 and across languages in Table 5.

We notice that our Legal-XLM-R-base model is on par with XLM-R large even though it only contains 33% of the parameters (184M vs 560M). All our models outperform XLM-R large on the dataset aggregate score. Our base model sets a new SotA on MAPA (MAP), the large model on CoViD 19 emergency event (C19) and the Longformer on Brazilian court decisions (BCD), German argument mining (GAM), Greek legal code (GLC) and Swiss judgment prediction (SJP). Surprisingly, the legal models slightly underperform in three NER tasks (GLN, LNR, and LNB). Sensitivity to hyperparameter choice could be a reason for this underperformance (we used the same hyperparameters for all models without tuning due to limited compute resources). We see the largest improvements over prior art in Brazilian court decisions (72.4 vs. 63.5) and in Greek legal code (70.2 vs 62.8). Maybe these tasks are particularly hard and therefore legal in-domain pretraining helps more. For BCD especially, the large amount of Brazilian caselaw in the pretraining corpus may offer an additional explanation.

The monolingual models underperform their base model XLM-R base only in Italian, Polish, and Romanian. In some languages the monolingual model even outperforms XLM-R base clearly (Estonian, Croatian, Hungarian, Lat-

Model	ECtHR-A	ECtHR-B	SCOTUS	EUR-LEX	LEDGAR	UNFAIR-ToS	CaseHOLD	Agg.
TFIDF+SVM *	48.9	63.8	64.4	47.9	81.4	75.0	22.4	49.0
BERT *	63.6	73.4	58.3	57.2	81.8	81.3	70.8	68.2
DeBERTa *	60.8	71.0	62.7	57.4	83.1	80.3	72.6	68.5
RoBERTa-base *	59.0	68.9	62.0	57.9	82.3	79.2	71.4	67.5
RoBERTa-large *	67.6	71.6	66.3	58.1	**83.6**	81.6	74.4	70.9
Longformer *	64.7	71.7	64.0	57.7	83.0	80.9	71.9	69.5
BigBird *	62.9	70.9	62.0	56.8	82.6	81.3	70.8	68.4
Legal-BERT *	64.0	74.7	66.5	57.4	83.0	**83.0**	75.3	70.8
CaseLaw-BERT *	62.9	70.3	65.9	56.6	83.0	82.3	75.4	69.7
Legal-en-R-base (ours)	65.2	73.7	66.4	**59.2**	82.7	78.7	73.3	70.5
Legal-en-R-large (ours)	**70.3**	**77.0**	**67.7**	58.4	82.5	82.4	**77.0**	**72.7**
Legal-XLM-R-base (ours)	64.8	73.9	63.9	58.2	82.8	79.6	71.7	69.7
Legal-XLM-R-large (ours)	68.2	74.2	67.5	58.4	82.7	79.9	75.1	71.4
Legal-XLM-LF-base (ours)	67.9	76.2	61.6	59.1	82.1	78.9	72.0	70.2

Table 6. Results on LexGLUE. We report macro-F1 and best scores in bold. Results from models marked with * are from Chalkidis et al. (2021d). Similar to LEXTREME, we computed the aggregate score as the harmonic mean of individual dataset results.

vian, Maltese, Dutch, Slovenian, and Swedish), and in five of them even set the new SotA for the language, sometimes clearly outperforming all other models (the Dutch model even outperforms its closest competitor mDeBERTa-v2 by 11.2 macro F1 and its base model XLM-R by almost 20 macro F1). These languages are all in the lower end of the data availability in the MULTILEGALPILE with the richest language (Dutch) containing only 810M words (see Figure 3). Pretraining a monolingual model on in-domain data may therefore be worth it, especially in low-resource languages.

Even though our legal Longformer model performs best on the dataset level, it performs much worse on the language level, possibly due to its lower scores in the most multilingual tasks MEU, MAP and C19 (24, 24 and 6 languages, respectively). Our legal base and large models achieve SotA in some languages, and are on aggregate almost as robust across languages as XLM-R.

Computing the final LEXTREME scores (harmonic mean of dataset aggregate and language aggregate scores), we find that the Legal-XLM-R-large is the new SotA on LEXTREME with a score of 59.5 vs 59.4 for Legal-XLM-R-base and 59.3 for XLM-R large. The legal Longformer's LEXTREME scores is with 56.5 not competitive due to its low language aggregate score.

5.4. Evaluation on LexGLUE

We evaluate our English and multilingual models on LexGLUE (Chalkidis et al., 2021e) and compare against baselines (see Table 6). Our models excel on the ECtHR, SCOTUS, EUR-LEX, and CaseHOLD tasks, achieving new SotA. In the other two tasks our models match general-purpose models such as RoBERTa. A reason for slight underperformance of the legal models in the LEDGAR and especially the Unfair ToS tasks may be the relatively low availability of contracts in the MULTILEGALPILE.

6. Conclusions and Future Work

Limitations We did not perform deduplication, thus data from the legal mC4 part might be present in other parts. However, recent work (Muennighoff et al., 2023) suggests that data duplication does not degrade performance during pretraining for up to four epochs. Overlap between the other parts is highly unlikely, since they are from completely different jurisdictions.

Conclusions Due to a general lack of multilingual pretraining data especially in specialized domains such as law, we curate a large-scale high-quality corpus in 24 languages from 17 jurisdictions. We continue pretraining XLM-R checkpoints on our data, achieving a new SotA for base and large models on the LEXTREME benchmark and vastly outperforming previous methods in greek legal code. We turn our XLM-R base model into a Longformer and continue pretraining on long documents. It reaches a new SotA in four LEXTREME datasets and reaches the overall highest dataset aggregate score. Monolingual models achieve huge gains over their base model XLM-R in some languages and even set language specific SotA in five languages outperforming other models by as much as 11 macro F1. On LexGLUE our English models reach SotA in five out of seven tasks with the large model achieving the highest aggregate score.

Future Work We leave the pretraining of a large generative multilingual legal language model for future work. Here we limited the corpus to the EU languages due to resource constraints, but in the future, we would like to expand the corpus in terms of languages and jurisdictions covered. Especially in China there exist many accessible

sources suitable to extend the corpus. Finally, it would be very interesting to study in more detail the specific contents of the MULTILEGALPILE.

References

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., and Lampos, V. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93, October 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.93. URL https://peerj.com/articles/cs-93. Publisher: PeerJ Inc.

Angelidis, I., Chalkidis, I., and Koubarakis, M. Named entity recognition, linking and generation for greek legislation. 2018.

Baisa, V., Michelfeit, J., Medveď, M., and Jakubíček, M. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2799–2803, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1445.

Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*, December 2020. URL http://arxiv.org/abs/2004.05150. arXiv: 2004.05150.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020b.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4317–4323, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1424. URL https://aclanthology.org/P19-1424.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4317–4323, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1424. URL https://www.aclweb.org/anthology/P19-1424.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261. URL https://aclanthology.org/2020.findings-emnlp.261.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]*, October 2020b. URL http://arxiv.org/abs/2010.02559. arXiv: 2010.02559.

Chalkidis, I., Fergadiotis, M., and Androutsopoulos, I. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *EMNLP*, 2021a.

Chalkidis, I., Fergadiotis, M., and Androutsopoulos, I. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv:2109.00904 [cs]*, September 2021b. URL http://arxiv.org/abs/2109.00904. arXiv: 2109.00904.

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., and Malakasiotis, P. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 226–241, Online, June 2021c. Association for Computational Linguistics.

doi: 10.18653/v1/2021.naacl-main.22. URL https://aclanthology.org/2021.naacl-main.22.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., and Aletras, N. Lexglue: A benchmark dataset for legal language understanding in english, 2021d. URL https://arxiv.org/abs/2110.00976.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M. J., Androutsopoulos, I., Katz, D. M., and Aletras, N. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. SSRN Scholarly Paper ID 3936759, Social Science Research Network, Rochester, NY, October 2021e. URL https://papers.ssrn.com/abstract=3936759.

Chalkidis*, I., Garneau*, N., Goanta, C., Katz, D. M., and Søgaard, A. Lexfiles and legallama: Facilitating english multinational legal language model development, 2023.

Conneau, A. and Lample, G. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*, April 2020b. URL http://arxiv.org/abs/1911.02116. arXiv: 1911.02116.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Drawzeski, K., Galassi, A., Jablonowska, A., Lagioia, F., Lippi, M., Micklitz, H. W., Sartor, G., Tagiuri, G., and Torroni, P. A Corpus for Multilingual Analysis of Online Terms of Service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 1–8, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.nllp-1.1.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027 [cs]*, December 2020a. URL http://arxiv.org/abs/2101.00027. arXiv: 2101.00027.

Gao, L., Biderman, S. R., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2020b.

Gokaslan, A. and Cohen, V. Openwebtext corpus, 2019. URL http://Skylion007.github.io/OpenWebTextCorpus.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3 (1), oct 2021. ISSN 2691-1957. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Gonzalez-Agirre, A., and Villegas, M. Spanish Legalese Language Model and Corpora. oct 2021. URL http://arxiv.org/abs/2110.12201.

He, P., Gao, J., and Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [cs]*, December 2021a. URL http://arxiv.org/abs/2111.09543. arXiv: 2111.09543.

He, P., Liu, X., Gao, J., and Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*, October 2021b. URL http://arxiv.org/abs/2006.03654. arXiv: 2006.03654.

Henderson, P., Krass, M. S., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. E. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, July 2022. URL http://arxiv.org/abs/2207.00220. arXiv:2207.00220 [cs].

Henderson*, P., Krass*, M. S., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. E. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset, 2022. URL https://arxiv.org/abs/2207.00220.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use, 2023.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Huang, K., Altosaar, J., and Ranganath, R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 2019. URL http://arxiv.org/abs/1904.05342.

ImaniGooghari, A., Lin, P., Kargaran, A. H., Severini, S., Sabet, M. J., Kassner, N., Ma, C., Schmid, H., Martins, A. F. T., Yvon, F., and Schütze, H. Glot500: Scaling multilingual corpora and language models to 500 languages, 2023.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. GPT-4 Passes the Bar Exam, March 2023. URL https://papers.ssrn.com/abstract=4389233.

Lage-Freitas, A., Allende-Cid, H., Santana, O., and Oliveira-Lage, L. Predicting Brazilian Court Decisions. *PeerJ Computer Science*, 8:e904, March 2022. ISSN 2376-5992. doi: 10.7717/peerj-cs.904. URL https://peerj.com/articles/cs-904. Publisher: PeerJ Inc.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, pp. btz682, September 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btz682. URL http://arxiv.org/abs/1901.08746. arXiv:1901.08746.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. ISSN 14602059. doi: 10.1093/bioinformatics/btz682.

Licari, D. and Comandè, G. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. Technical report, 2022. URL http://ceur-ws.org.

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., and Torroni, P. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139, 2019. ISSN 1572-8382. doi: 10.1007/s10506-019-09243-2. URL https://doi.org/10.1007/s10506-019-09243-2.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. 2019. URL http://arxiv.org/abs/1907.11692.

Luz de Araujo, P. H., Campos, T. E. d., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pp. 313–323. Springer, 2018. Dataset URL: https://huggingface.co/datasets/lener_br.

Masala, M., Iacob, R. C. A., Uban, A. S., Cidota, M., Velicu, H., Rebedea, T., and Popescu, M. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 86–94, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nllp-1.8. URL https://aclanthology.org/2021.nllp-1.8.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models, 2016.

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., and Raffel, C. Scaling Data-Constrained Language Models, May 2023. URL http://arxiv.org/abs/2305.16264. arXiv:2305.16264 [cs].

Niklaus, J. and Giofré, D. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch?, November 2022. URL http://arxiv.org/abs/2211.17135. arXiv:2211.17135 [cs].

Niklaus, J., Chalkidis, I., and Stürmer, M. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 19–35, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.nllp-1.3.

Niklaus, J., Stürmer, M., and Chalkidis, I. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 32–46, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.3.

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., and Chalkidis, I. Lextreme: A multi-lingual and multi-task benchmark for the legal domain, 2023. URL https://arxiv.org/abs/2301.13126.

OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Pais, V., Mitrofan, M., Gasan, C. L., Coneschi, V., and Ianov, A. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 9–18, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nllp-1.2. URL https://aclanthology.org/2021.nllp-1.2.

Papaloukas, C., Chalkidis, I., Athinaios, K., Pantazi, D.-A., and Koubarakis, M. Multi-granular legal topic classification on greek legislation. *arXiv preprint arXiv:2109.15298*, 2021. Dataset URL: https://huggingface.co/datasets/greek_legal_code.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL https://aclanthology.org/2021.emnlp-main.800.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020b. ISSN 1533-7928. URL http://jmlr.org/papers/v21/20-074.html.

Sag, M. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*, 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, February 2020. URL http://arxiv.org/abs/1910.01108. arXiv: 1910.01108.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020.

Spaeth, H. J., Epstein, L., Martin, A. D., Segal, J. A., Ruger, T. J., and Benesh, S. C. Supreme Court Database, Version 2020 Release 01.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A Large Language Model for Science, November 2022. URL http://arxiv.org/abs/2211.09085. arXiv:2211.09085 [cs, stat].

Tiedemann, J. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3518–3522, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1559.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E.,

Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.1, 2023.

Tuggener, D., von Däniken, P., Peetz, T., and Cieliebak, M. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1235–1241, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.155.

Tziafas, G., de Saint-Phalle, E., de Vries, W., Egger, C., and Caselli, T. A multilingual approach to identify and classify exceptional measures against covid-19. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 46–62, 2021. Dataset URL: https://tinyurl.com/ycysvtbm.

Urchs, S., Mitrović, J., and Granitzer, M. Design and Implementation of German Legal Decision Corpora:. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pp. 515–521, Online Streaming, — Select a Country —, 2021. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-484-8. doi: 10.5220/0010187305150521. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010187305150521.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. pp. 30, 2019.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL https://arxiv.org/abs/2109.01652.

Wettig, A., Gao, T., Zhong, Z., and Chen, D. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2985–3000, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.217.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*, March 2021. URL http://arxiv.org/abs/2010.11934. arXiv: 2010.11934.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. *Defending against Neural Fake News*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pp. 159–168, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385268. doi: 10.1145/3462757.3466088. URL https://doi.org/10.1145/3462757.3466088.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015.

Model Name	# Steps	Vocab Size
Legal-bg-R-base	200K	32K
Legal-hr-R-base	200K	32K
Legal-cs-R-base	200K	32K
Legal-da-R-base	200K	32K
Legal-nl-R-base	200K	32K
Legal-en-R-base	200K	32K
Legal-en-R-large	500K	32K
Legal-et-R-base	200K	32K
Legal-fi-R-base	200K	32K
Legal-fr-R-base	200K	32K
Legal-de-R-base	200K	32K
Legal-el-R-base	200K	32K
Legal-hu-R-base	200K	32K
Legal-ga-R-base	200K	32K
Legal-it-R-base	200K	32K
Legal-lv-R-base	200K	32K
Legal-lt-R-base	200K	32K
Legal-mt-R-base	200K	32K
Legal-pl-R-base	200K	32K
Legal-pt-R-base	200K	32K
Legal-ro-R-base	200K	32K
Legal-sk-R-base	200K	32K
Legal-sl-R-base	200K	32K
Legal-es-R-base	200K	32K
Legal-sv-R-base	200K	32K
Legal-XLM-R-base	1M	128K
Legal-XLM-R-large	500K	128K
Legal-XLM-LF-base	50K	128K

Table 7. Model Details

A. Training Details

B. Hyperparameter Details

source	Dataset	Task	Task type	Hierarchical	Seeds	lower case	Batch size	Metric for best model	Evaluation strategy	Epochs	Early stopping patience	Learning rate
(Niklaus et al., 2023)	GLN	GLN	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	LNR	LNR	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	LNB	LNB	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	MAP	MAP-F	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	MAP	MAP-C	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	BCD	BCD-J	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	BCD	BCD-U	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GAM	GAM	SLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GLC	GLC-C	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GLC	GLC-S	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GLC	GLC-V	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	SJP	SJP	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	OTS	OTS-UL	SLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	OTS	OTS-CT	MLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	C19	C19	MLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	MEU	MEU-1	MLTC	True	1,2,3	True	64	evaluation loss			5	1e-5
(Niklaus et al., 2023)	MEU	MEU-2	MLTC	True	1,2,3	True	64	evaluation loss			5	1e-5
(Niklaus et al., 2023)	MEU	MEU-3	MLTC	True	1,2,3	True	64	evaluation loss			5	1e-5
(Chalkidis et al., 2021d)	ECtHR	ECtHR-A	MLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	ECtHR	ECtHR-B	MLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	EUR-LEX	EUR-LEX	MLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	SCOTUS	SCOTUS	SLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	LEDGAR	LEDGAR	SLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	UnfairToS	UnfairToS	MLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	CaseHOLD	CaseHOLD	MCQA	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5

Table 8. Hyperparameters for each dataset and task. However, there were a few exceptions. For the multilingual MEU tasks, given the dataset's size, we trained them for only 1 epoch with 1000 steps as the evaluation strategy when using multilingual models. When using monolingual models, we trained for 50 epochs with epoch-based evaluation strategy, as we utilized only the language-specific subset of the dataset. Regarding LexGlue, we followed the guidelines of Chalkidis et al. (2021d) for RoBERTa-based large language models, which required a maximum learning rate of 1e-5, a warm-up ratio of 0.1, and a weight decay rate of 0.06.

C. Dataset Details

Language	Text Type	Words	Documents	Words per Document	Jurisdiction	Source	License/Copyright
Native Multi Legal Pile							
bg	legislation	309M	262k	1178	Bulgaria	MARCELL	CC0-1.0
cs	caselaw	571M	342k	1667	Czechia Czechia Czechia	CzCDC Constitutional Court CzCDC Supreme Administrative Court CzCDC Supreme Court	CC BY-NC 4.0 CC BY-NC 4.0 CC BY-NC 4.0
da	caselaw	211M	92k	2275	Denmark	DDSC	CC BY 4.0 and other, depending on the dataset
da	legislation	653M	296k	2201	Denmark	DDSC	CC BY 4.0 and other, depending on the dataset
de	caselaw	1786M	614k	2905	Germany Switzerland	openlegaldata entscheidsuche	ODbL-1.0 similar to CC BY
de	legislation	513M	302k	1698	Germany Switzerland	openlegaldata lexfind	ODbL-1.0 not protected by copyright law
en	legislation	2539M	713k	3557	Switzerland UK	lexfind uk-lex	not protected by copyright law CC BY 4.0
fr	caselaw	1172M	495k	2363	Belgium France Luxembourg Switzerland	jurportal CASS judoc entscheidsuche	not protected by copyright law Open Licence 2.0 not protected by copyright law similar to CC BY
fr	legislation	600M	253k	2365	Switzerland Belgium	lexfind ejustice	not protected by copyright law not protected by copyright law
hu	legislation	265M	259k	1019	Hungary	MARCELL	CC0-1.0
it	caselaw	407M	159k	2554	Switzerland	entscheidsuche	similar to CC BY
it	legislation	543M	238k	2278	Switzerland	lexfind	not protected by copyright law
nl	legislation	551M	243k	2263	Belgium	ejustice	not protected by copyright law
pl	legislation	299M	260k	1148	Poland	MARCELL	CC0-1.0
pt	caselaw	12613M	17M	728	Brazil Brazil Brazil	RulingBR CRETA CJPG	not protected by copyright law CC BY-NC-SA 4.0 not protected by copyright law
ro	legislation	559M	396k	1410	Romania	MARCELL	CC0-1.0
sk	legislation	280M	246k	1137	Slovakia	MARCELL	CC0-1.0
sl	legislation	366M	257k	1418	Slovenia	MARCELL	CC-BY-4.0
total		**24236M**	**23M**	**1065**		**Native Multi Legal Pile**	
Overall statistics for the remaining subsets							
total		**12107M**	**8M**	**1457**	EU	**Eurlex Resources**	**CC BY 4.0**
total		**43376M**	**18M**	**2454**	US (99%), Canada, and EU	**Pile of Law**	CC BY-NC-SA 4.0; **See Henderson* et al. (2022) for details**
total		**28599M**	**10M**	**2454**		**Legal mC4**	**ODC-BY**

Table 9. Information about size and number of words and documents for *Native* Multi Legal Pile are provided according to language and text type. For the remaining subsets of Multi Legal Pile we provide general statistics.