MultiLegalPile: A 689GB Multilingual Legal Corpus

Joel Niklaus[1,2,4] **Veton Matoshi**[2]

Matthias Stürmer[1,2] **Ilias Chalkidis**[3] **Daniel E. Ho**[4]

[1]University of Bern [2]Bern University of Applied Sciences
[3]University of Copenhagen [4]Stanford University

Abstract

Large, high-quality datasets are crucial for training Large Language Models (LLMs). However, so far, few datasets are available for specialized critical domains such as law and the available ones are often small and only in English. To fill this gap, we curate and release MULTILEGALPILE, a 689GB corpus in 24 languages from 17 jurisdictions. MULTILEGALPILE includes diverse legal data sources and allows for pretraining NLP models under fair use, with most of the dataset licensed very permissively. 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, trained models, and all code under the most open licenses possible.

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 practice 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., 2020), and mC4 (Xue et al., 2021). However, 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[1], 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. We find that only one in every thousand document in mC4 contains legal text. Similarly, LLMs are predominantly English, especially considering domain-specific models, e.g., ones specialized in biomedical, legal, or financial texts.

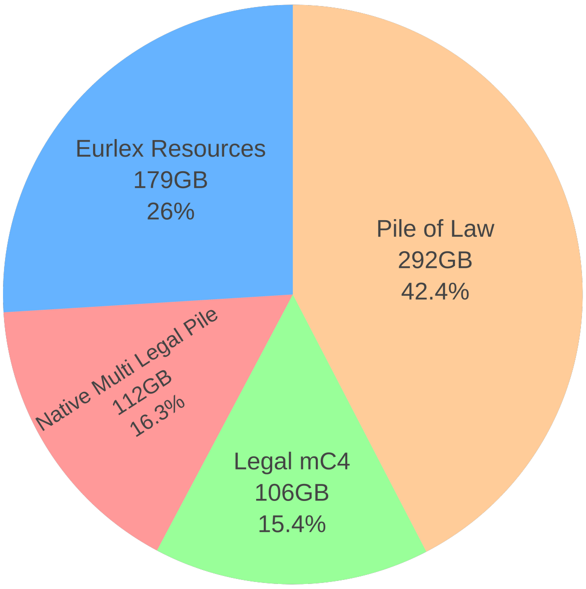

Figure 1: MULTILEGALPILE Source Distribution

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 massive, 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 and Lample, 2019) on our corpus and evaluated these models on the recently introduced LEXTREME (Niklaus et al., 2023) and LexGLUE (Chalkidis et al., 2022) benchmarks. Given the often extensive nature of legal text, we also pretrained a Longformer model (Beltagy et al.,

[1] https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15077–15094
August 11-16, 2024 ©2024 Association for Computational Linguistics

2020) for comparison with hierarchical models (Chalkidis et al., 2019; Niklaus et al., 2021, 2022).

Our multilingual models set a new state-of-the-art (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 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 LLMs.

Contributions

The contributions of this paper are three-fold: First, we curate and release a large multilingual legal text corpus, dubbed MULTILEGALPILE,[2] covering 24 languages and 17 legal systems (jurisdictions). Second, we release two multilingual and 24 monolingual legal PLMs, termed LEGALXLMS, initiated from XLM-R (Conneau and Lample, 2019) and further pretrained on the MULTILEGALPILE. We also pretrain a Longformer (Beltagy et al., 2020) based on our multilingual base-size model on context lengths of up to 4096 tokens. Third, we benchmark the newly released models on LEXTREME and LexGLUE, reaching SotA for base- and large-size models and increasing performance drastically in Greek legal code. Our Longformer model achieves SotA in four tasks and the highest dataset aggregate score. Our monolingual models set language-specific SotA in five languages.

2 Related Work

In this section, we briefly discuss prior general and domain-specific pretraining corpora. See Appendix B for a more detailed discussion of related works.

2.1 General Pretraining Corpora

The One Billion Word Language Model Benchmark (LM1B) (Chelba et al., 2014), Wikipedia, and derived datasets like WikiText (Merity et al., 2016) and BookCorpus (Zhu et al., 2015) have been crucial in developing language models such as GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). Large-scale datasets like the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020), OpenWebText (Gokaslan and Cohen, 2019), The Pile (Gao et al., 2020b), and Glot500 (ImaniGooghari et al., 2023) have further advanced the field, contributing to the training of models like T5, MegatronBERT (Shoeybi et al., 2020), GPT-3 (Brown et al., 2020), and Glot500-m. Although general pretraining datasets are large and widely available, we find that mC4 only contains around 0.1% legal text (see Section 3.1), exemplifying the need for datasets specifically tailored to the legal domain.

2.2 Domain Specific 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., 2020)	biomedical	English	18B
LegalBERT (Chalkidis et al., 2020)	legal	English	1.44B (11.5GB)
CaselawBERT (Zheng et al., 2021)	legal	English	4.63B (37GB)
LexFiles (Chalkidis et al., 2020)	legal	English	18.8B
LegalXLMs (ours)	legal	24 EU langs	87B (689GB)

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

Pretraining on domain-specific text like medicine, law, or science can boost Language Model (LM) performance on related tasks (Beltagy et al., 2019; Gu et al., 2021; Chalkidis et al., 2020; Niklaus and Giofré, 2022). In the scientific field, SciBERT was pretrained on a mix of computer science and biomedical papers (Beltagy et al., 2019). Similarly, models like PubMedBERT (Gu et al., 2021) and BioBERT (Lee et al., 2020) were pretrained on biomedical datasets. ClinicalBERT utilized the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, encompassing 2 million clinical notes, demonstrating superior performance on medical NLP tasks (Huang et al., 2019). In the legal realm, LegalBERT was pretrained on 12 GB of English legal texts, achieving high performance on domain-specific tasks (Chalkidis et al., 2020). CaseLaw-BERT utilized the English Harvard Law case corpus from 1965 to 2021 (Zheng et al., 2021). Recently, LexFiles was released, with 11 sub-corpora covering six English-speaking legal systems and 19B tokens (Chalkidis* et al., 2023). It was used to train new legal English PLMs, showing enhanced results in legal tasks. Though efforts to

[2] Link will be released upon acceptance.

pretrain legal LMs exist in languages like Italian, Romanian, and Spanish (Licari and Comandè, 2022; Masala et al., 2021; Gutiérrez-Fandiño et al., 2021), English remains predominant, emphasizing the need for multilingual legal corpora. Table 1 compares previous domain-specific corpora, all in English and all legal corpora less than 1/4 of MULTILEGALPILE's size.

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 used the vast multilingual web crawl corpus, mC4 (Xue et al., 2021), as our base dataset. To effectively isolate legal content, we used regular expressions to identify documents with legal references, such as "Art. 5" or "§ 8" . 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.

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

Table 2: Per language precision in legal mC4 (n=20)

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, 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 using regexes instead of an machine learning 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 on our hardware. 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.

We find that on average, only one in every thousand pages in mC4 contains legal content. We show a precise overview of language-specific percentages of legal text in mC4 in Figure 4.

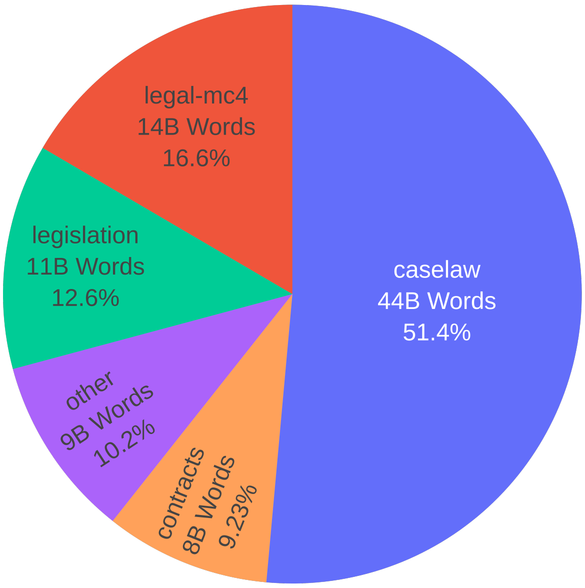

Figure 2: MULTILEGALPILE Text Type Distribution

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. Most sources offered direct data download links. For inconsistently formatted data, we converted them to a unified format like jsonl. The post-processing steps involved performing various tasks depending on the initial data format. For example, in the case of CASS[3], we extracted the textual data from XML tags.

Curating Eurlex Resources To curate the Eurlex resources, we utilized the eurlex R package (Ovádek, 2021) to generate SPARQL queries and

[3] https://echanges.dila.gouv.fr/OPENDATA/CASS

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 subsets: a) Native Multi Legal Pile (112 GB), b) Eurlex Resources (179 GB), c) Legal MC4 (106 GB) and d) Pile of Law (Henderson et al., 2022) (292 GB). Figure 3 details the distribution of languages. Due to Pile of Law integration, English dominates, comprising over half the words. Figure 2 shows text type distribution. Caselaw comprises over half the corpus, due to the good public access to court rulings especially in common law countries. Even in civil law countries, where legislation is crucial, caselaw often outnumbers legislation, as seen in the Swiss case in Table 9. Publicly available contracts are scarce, contributing less than 10% to the corpus, despite potentially making up most existing legal texts (from the private sector). Note that most contracts in our corpus originate from the US or EU international treaties. Table 9 in Appendix E provides additional information on 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 to the 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 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 CC BY 4.0 licensed by the European Union[4], thus posing no 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[5] 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 unaltered, thus preserving the license.

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 multilingual 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 the LM development literature:

(a) We warm-start (initialize) our models from the original XLM-R checkpoints (base or large) of Conneau and 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. Similarly to Liang et al. (2023) who increased the vocabulary size from around 2.5K tokens per language (250K for 100 languages) to around 10K (1M for 100 languages),

[4] https://eur-lex.europa.eu/content/legal-notice/legal-notice.html

[5] https://huggingface.co/datasets/mc4

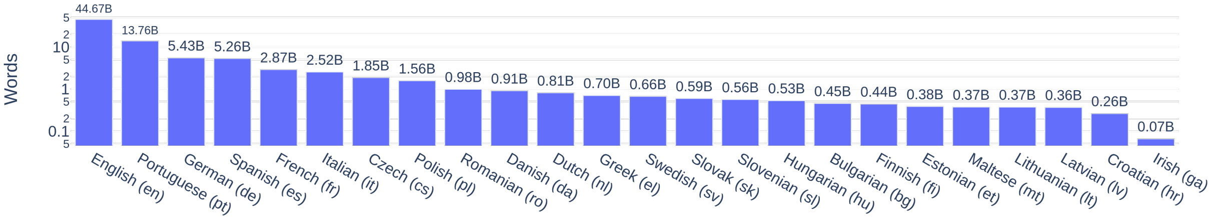

Figure 3: MULTILEGALPILE Language Distribution (Note the log-scaled y-axis)

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. (2020)	278M	250K	1.5M steps / BS 8192	2.5TB CC100	100
XLM-R large	Conneau et al. (2020)	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 process up to 512 tokens, except Legal-XLM-LF-base (4096 tokens). BS is short for batch size. MLP is short for MULTILEGALPILE. Params is the total parameter count (including embedding layer).

we increased to around 5K (128K for 24 languages), thus roughly doubling compared to XLM-R.

(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 100% of token predictions are based on masked tokens, compared to Devlin et al. (2019)[6], 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 and Lample (2019) and Raffel et al. (2020), 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 recently developed large PLMs (Conneau and Lample, 2019; Raffel et al., 2020; Brown et al., 2020).

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 $3e$–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 C 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.[7]

[6] 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.

[7] Link will be released upon acceptance.

5 Evaluating on Legal Benchmarks

In the absence of established legal benchmarks for generative tasks and our focus on pretraining encoder-only models, we select two established legal benchmarks involving challenging text classification and named entity recognition tasks involving long documents: LEXTREME and LexGLUE.

5.1 Benchmark Description

Below, we briefly describe each dataset and refer the reader to the original works 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) contains 200 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., 2021a) 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., 2022) is a legal benchmark covering two single-label and, four multi-label text classification datasets, and a multiple choice question answering dataset.

ECtHR Tasks A & B (Chalkidis et al., 2019, 2021b) contain approx. 11K cases from the European Court of Human Rights (ECtHR) public database. Based on case facts, Task A predicts violated articles and Task B allegedly violated articles of the European Convention of Human Rights (ECHR). **SCOTUS** (Spaeth et al., 2020) 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 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 out 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., 2022) 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. We follow previous work (Niklaus et al., 2023; Chalkidis et al., 2022) and do not tune hyperparameters.

All scores are macro-F1 scores, equally weighing each class for fairness in unbalanced datasets. To obtain Table 6, we follow Chalkidis et al. (2022), running five repetitions with different random seeds (1-5) and report test scores from the best-performing seed on the development data. For values in Tables 4 and 5, we follow the procedure in Niklaus et al. (2023), taking the harmonic mean of the results of 3 random seeds (1-3). We calculate

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 (macro-F1) for multilingual models on LEXTREME with 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 (macro-F1) for multilingual models on LEXTREME with the best scores in **bold**. For each language, we list the top-performing monolingual legal and non-legal models under *NativeLegalBERT* and *NativeBERT*, and our legal models under *Legal-mono-R-base*. Missing values signify no suitable models found.

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 (macro-F1) with the best scores in **bold**. Results marked with * are from Chalkidis et al. (2022). Similar to LEXTREME, we calculate the aggregate score as the harmonic mean of dataset results.

the dataset aggregate in Table 4 by successively taking the harmonic mean of (i) the languages in the configurations (e.g., de,fr,it in SJP), (ii) configurations within datasets (e.g., OTS-UL, OTS-CT in OTS), and (iii) datasets in LEXTREME (BCD, GAM). The language aggregate score in Table 5 is computed similarly: by taking the harmonic mean of (i) configurations within datasets, (ii) datasets for each language (e.g., MAP, MEU for lv), and (iii) languages in LEXTREME (bg,cs). We show an overview of the models evaluated in Table 3.

Note that most LLMs are predominantly trained on English and Chinese with the exception of mT5 (Xue et al., 2021) and BLOOM (Scao et al., 2022) (more than 95% of LLaMA's pretraining corpus is English (Touvron et al., 2023)). Because LEXTREME and LexGLUE consist of NLU tasks, we compare to encoder-only LMs only.

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 BCD (72.4 vs. 63.5) and in GLC (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, Latvian, 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 in 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 score (56.5) is 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., 2022) and compare with baselines (see Table 6). Our models excel on the ECtHR, SCOTUS, EUR-LEX, and CaseHOLD tasks, setting 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

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. To conclude, following best practices in continued pretraining on our comprehensive multilingual legal corpus establishes new state-of-the-art across multiple datasets and languages, significantly enhancing performance in legal text analysis.

Future Work We focused on the 24 EU languages, 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. Additionally, we would like to find out whether our findings on in-domain pretraining hold for multi-billion generative models. Finally, a detailed examination of the contents of the MULTILEGALPILE could provide valuable insights into its structure and efficacy in enhancing legal language models.

Ethics Statement

This study focuses on evaluating legal-specific LMs from multiple aspects, expanding the dialogue to aid in creating support technologies for both legal professionals and the general public. This area represents a vital field for research, as emphasized by Tsarapatsanis and Aletras (2021), aiming to enhance legal services and make legal knowledge more accessible. The study also aims to shed light on the multifaceted limitations that need addressing to ensure the responsible and ethical application of legal-oriented technologies.

In pursuit of these goals, we introduce novel resources encompassing a range of legal systems. These resources are designed to construct new models that more accurately reflect legal nuances and evaluate their effectiveness more precisely. All resources created and shared in this work are derived from data that is publicly accessible, often distributed across various online platforms.

Limitations

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

Due to limited compute, we were not able to pretrain a large generative model and leave this to future work.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

I. Angelidis, Ilias Chalkidis, and M. Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *JURIX*.

Vít Baisa, Jan Michelfeit, Marek Medved', and Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2799–2803, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*. ArXiv: 2004.05150.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021b. 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*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A Benchmark

Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A Corpus for Multilingual Analysis of Online Terms of Service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 1–8, Punta Cana, Dominican Republic. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020a. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027 [cs]*. ArXiv: 2101.00027.

Leo Gao, Stella Rose Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020b. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1).

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish Legalese Language Model and Corpora. ArXiv:2110.12201 [cs].

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [cs]*. ArXiv: 2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*. ArXiv: 2006.03654.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. ArXiv:2207.00220 [cs].

Peter Henderson, Mark Simon Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel E Ho. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation models and fair use.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. ArXiv:2009.03300 [cs].

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. GPT-4 Passes the Bar Exam.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting Brazilian Court Decisions. *PeerJ Computer Science*, 8:e904. Publisher: PeerJ Inc.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models. ArXiv:2301.10472 [cs].

Daniele Licari and Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

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:1907.11692 [cs]*. ArXiv: 1907.11692.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer. Dataset URL: https://huggingface.co/datasets/lener_br.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling Data-Constrained Language Models. ArXiv:2305.16264 [cs].

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? ArXiv:2211.17135 [cs].

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. ArXiv:2301.13126 [cs].

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. 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)*, pages 32–46, Online only. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Michal Ovádek. 2021. Facilitating access to data on european union laws. *Political Research Exchange*, 3(1):1870150.

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on greek legislation. *arXiv preprint arXiv:2109.15298*. Dataset URL: https://huggingface.co/datasets/greek_legal_code.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

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

Matthew Sag. 2023. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko

Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. ArXiv:2211.05100 [cs].

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism.

Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2020. Supreme Court Database, Version 2020 Release 01.

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. ArXiv:2211.09085 [cs, stat].

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the Ethical Limits of Natural Language Processing on Legal Text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. 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*, pages 1235–1241, Marseille, France. European Language Resources Association.

Georgios Tziafas, Eugenie de Saint-Phalle, Wietse de Vries, Clara Egger, and Tommaso Caselli. 2021. A multilingual approach to identify and classify exceptional measures against covid-19. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 46–62. Dataset URL: https://tinyurl.com/ycysvtbm.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2021. Design and Implementation of German Legal Decision Corpora:. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pages 515–521, Online Streaming, — Select a Country —. SCITEPRESS - Science and Technology Publications.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. page 30.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*. ArXiv: 2010.11934.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. *Defending against Neural Fake News*. Curran Associates Inc., Red Hook, NY, USA.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. 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, pages 159–168, New York, NY, USA. Association for Computing Machinery.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies:

Towards story-like visual explanations by watching movies and reading books.

A Use of AI assistants

We used ChatGPT and Grammarly for improving the grammar and style of our writing.

B Additional Related Work

B.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.

One of the earliest widely-used datasets is the One Billion Word Language Model Benchmark (LM1B) (Chelba et al., 2014). It was created by extracting one billion words from web pages to evaluate novel language modeling techniques. It has been used, among others, to evaluate GPT-2 (Radford et al., 2019).

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., 2020), and GPT-3 (Brown et al., 2020), 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., 2020). 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., 2020) as well as XLM-R (Conneau et al., 2020). 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., 2020). 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., 2020) 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., 2020) and Switch Transformer (Fedus et al., 2022).

OpenWebText (Gokaslan and 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[8], 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 and Cohen, 2019), OpenSubtitles (Tiedemann, 2016) etc., encompassing 825 GB of data. Besides general-purpose textual datasets, it also 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

[8] https://commoncrawl.org/2016/10/news-dataset-available

et al., 2020) were evaluated on this dataset.

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[9] 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.

B.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., 2020; Niklaus and 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. “a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital”. 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. (2020) collected 12 GB of diverse English legal texts, including legislation, court cases, and contracts. They pretrained Legal-BERT on this dataset, showing SotA 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 Lex-Files, 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 and 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.

C Training Details

For finetuning the pretrained models on the evaluation benchmarks we used the following NVIDIA GPUs: 24GB RTX3090, 32GB V100 and 80GB A100. We used v3-8 TPUs for pretraining. All our experiments were run on Linux machines (Debian).

[9] https://github.com/togethercomputer/RedPajama-Data

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

D 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., 2022)	ECtHR	ECtHR-A	MLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2022)	ECtHR	ECtHR-B	MLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2022)	EUR-LEX	EUR-LEX	MLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2022)	SCOTUS	SCOTUS	SLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2022)	LEDGAR	LEDGAR	SLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2022)	UnfairToS	UnfairToS	MLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2022)	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. (2022) 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.

E 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. 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.

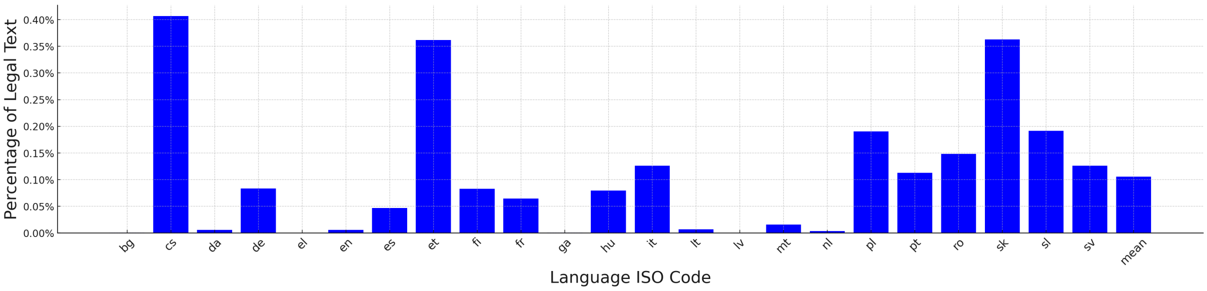

Figure 4: Percentage of Legal Text in mC4 per Language