LexGLUE
A Benchmark Dataset for Legal Language Understanding in English
Chalkidis, Ilias; Jana, Abhik; Hartung, Dirk; Bommarito, Michael; Androutsopoulos, Ion; Katz, Daniel Martin; Aletras, Nikolaos

Published in:
ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
LexGLUE: A Benchmark Dataset for Legal Language Understanding in English

Ilias Chalkidis α∗ Abhik Jana β Dirk Hartung γ δ Michael Bommarito γ δ
Ion Androutsopoulos e Daniel Martin Katz γ δ ζ Nikolaos Aletras η

α University of Copenhagen, Denmark β Universit¨at Hamburg, Germany
γ Bucerius Law School, Hamburg, Germany δ CodeX, Stanford Law School, United States
ε Athens University of Economics and Business, Greece a University of Sheffield, UK
ζ Illinois Tech – Chicago Kent College of Law, United States

Abstract

Laws and their interpretations, legal arguments and agreements are typically expressed in writing, leading to the production of vast corpora of legal text. Their analysis, which is at the center of legal practice, becomes increasingly elaborate as these collections grow in size. Natural language understanding (NLU) technologies can be a valuable tool to support legal practitioners in these endeavors. Their usefulness, however, largely depends on whether current state-of-the-art models can generalize across various tasks in the legal domain. To answer this currently open question, we introduce the Legal General Language Understanding Evaluation (LexGLUE) benchmark, a collection of datasets for evaluating model performance across a diverse set of legal NLU tasks in a standardized way. We also provide an evaluation and analysis of several generic and legal-oriented models demonstrating that the latter consistently offer performance improvements across multiple tasks.

1 Introduction

Law is a field of human endeavor dominated by the use of language. As part of their professional training, law students consume large bodies of text as they seek to tune their understanding of the law and its application to help manage human behavior. Virtually every modern legal system produces massive volumes of textual data (Katz et al., 2020). Lawyers, judges, and regulators continuously author legal documents such as briefs, memos, statutes, regulations, contracts, patents and judicial decisions (Coupette et al., 2021). Beyond the consumption and production of language, law and the art of lawyering is also an exercise centered around the analysis and interpretation of text.

Natural language understanding (NLU) technologies can assist legal practitioners in a variety of legal tasks (Chalkidis and Kampas, 2018; Aletras et al., 2019, 2020; Zhong et al., 2020b; Bommarito et al., 2021), from judgment prediction (Aletras et al., 2016; Sim et al., 2016; Katz et al., 2017; Zhong et al., 2018; Chalkidis et al., 2019a; Malik et al., 2021), information extraction from legal documents (Chalkidis et al., 2018, 2019c; Chen et al., 2020; Hendrycks et al., 2021) and case summarization (Bhattacharya et al., 2019) to legal question answering (Ravichander et al., 2019; Kien et al., 2020; Zhong et al., 2020a,c) and text classification (Nal- lapati and Manning, 2008; Chalkidis et al., 2019b, 2020a). Transformer models (Vaswani et al., 2017) pre-trained on legal, rather than generic, corpora have also been studied (Chalkidis et al., 2020b; Zheng et al., 2021; Xiao et al., 2021).

Pre-trained Transformers, including BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), BART (Lewis et al., 2020), DeBERTa (He et al., 2021) and numerous variants, are currently the state of the art in most natural language processing (NLP) tasks. Rapid performance improvements have been witnessed, to the extent that ambitious multi-task benchmarks (Wang et al., 2018, 2019b) are considered almost ‘solved’ a few years after their release and need to be made more challenging (Wang et al., 2019a).
Recently, Bommasani et al. (2021) named these pre-trained models (e.g., BERT, DALL-E, GPT-3) foundation models. The term may be controversial, but it emphasizes the paradigm shift these models have caused and their interdisciplinary potential. Studying the latter includes the question of how to adapt these models to legal text (Bommarito et al., 2021). As discussed by Zhong et al. (2020b) and Chalkidis et al. (2020b), legal text has distinct characteristics, such as terms that are uncommon in generic corpora (e.g., 'restrictive covenant', 'promissory estoppel', 'tort', 'novation'), terms that have different meanings than in everyday language (e.g., an 'executed' contract is signed and effective, a 'party' is a legal entity), older expressions (e.g., pronominal adverbs like 'herein', 'hereto', 'wherefore'), uncommon expressions from other languages (e.g., 'laches', 'voir dire', 'certiorari', 'sub judice'), and long sentences with unusual word order (e.g., "the provisions for termination hereinafter appearing or will at the cost of the borrower forthwith comply with the same") to the extent that legal language is often classified as a 'sub-language' (Tiersma, 1999; Williams, 2007; Haigh, 2018). Furthermore, legal documents are often much longer than the maximum length state-of-the-art deep learning models can handle, including those designed to handle long text (Beltagy et al., 2020; Zaheer et al., 2020; Yang et al., 2020).

Inspired by the recent widespread use of the GLUE multi-task benchmark NLP dataset (Wang et al., 2018, 2019b), the subsequent more difficult SuperGLUE (Wang et al., 2019a), other previous multi-task NLP benchmarks (Conneau and Kiela, 2018; McCann et al., 2018), and similar initiatives in other domains (Peng et al., 2019), we introduce LexGLUE, a benchmark dataset to evaluate the performance of NLP methods in legal tasks. LexGLUE is based on seven English existing legal NLP datasets, selected using criteria largely from SuperGLUE (discussed in Section 3.1). As in GLUE and SuperGLUE (Wang et al., 2018, 2019b,a), one of our goals is to push towards generic (or ‘foundation’) models that can cope with multiple NLP tasks, in our case legal NLP tasks, possibly with limited task-specific fine-tuning. Another goal is to provide a convenient and informative entry point for NLP researchers and practitioners wishing to explore or develop methods for legal NLP. Having these goals in mind, the datasets we include in LexGLUE and the tasks they address have been simplified in several ways, discussed below, to make it easier for newcomers and generic models to address all tasks. We provide Python APIs integrated with Hugging Face (Wolf et al., 2020; Lhoest et al., 2021) to easily import all the datasets we experiment with and evaluate the performance of different models (Section 4.4).

By unifying and facilitating the access to a set of law-related datasets and tasks, we hope to attract not only more NLP experts, but also more interdisciplinary researchers (e.g., law doctoral students willing to take NLP courses). More broadly, we hope LexGLUE will speed up the adoption and transparent evaluation of new legal NLP methods and approaches in the commercial sector, too. Indeed, there have been many commercial press releases in the legal tech industry on high-performing systems, but almost no independent evaluation of the performance of machine learning and NLP-based tools. A standard publicly available benchmark would also allay concerns of undue influence in predictive models, including the use of metadata which the relevant law expressly disregards.

## 2 Related Work

The rapid growth of the legal text processing field is demonstrated by numerous papers presented in top-tier conferences in NLP and artificial intelligence (Luo et al., 2017; Zhong et al., 2018; Chalkidis et al., 2019a; Valvoda et al., 2021) as well as surveys (Chalkidis and Kampas, 2018; Zhong et al., 2020b; Bommarito et al., 2021). Moreover, specialized workshops on NLP for legal text (Aletras et al., 2019; Di Fatta et al., 2020; Aletras et al., 2020) are regularly organized.

A core task in this area has been legal judgment prediction (forecasting), where the goal is to predict the outcome (verdict) of a court case. In this direction, there have been at least three lines of work. The first one (Aletras et al., 2016; Chalkidis et al., 2019a; Medvedeva et al., 2020, 2021) predicts violations of human rights in cases of the
European Court of Human Rights (ECtHR). The second line of work (Luo et al., 2017; Zhong et al., 2018; Yang et al., 2019) considers Chinese criminal cases where the goal is to predict relevant law articles, criminal charges, and the term of the penalty. The third line of work (Ruger et al., 2004; Katz et al., 2017; Kaufman et al., 2019) includes methods for predicting the outcomes of cases of the Supreme Court of the United States (SCOTUS).

The same or similar tasks have also been studied with court cases in many other jurisdictions including France (Şulea et al., 2017), Philippines (Virtucio et al., 2018), Turkey (Mumcuoğlu et al., 2021), Thailand (Kowsrihawat et al., 2018), United Kingdom (Strickson and De La Iglesia, 2020), Germany (Urchs et al., 2021), and Switzerland (Niklaus et al., 2021). Apart from predicting court decisions, there is also work aiming to interpret (explain) the decisions of particular courts (Ye et al., 2018; Chalkidis et al., 2021c; Branting et al., 2021).

Another popular task is legal topic classification. Nallapati and Manning (2008) highlighted the challenges of legal document classification compared to more generic text classification by using a dataset including docket entries of US court cases. Chalkidis et al. (2020a) classify EU laws into EuroVoc concepts, a task earlier introduced by Mencia and Fürnkranzand (2007), with a special interest in few- and zero-shot learning. Luz de Araujo et al. (2020) also studied topic classification using a dataset of Brazilian Supreme Court cases. There are similar interesting applications in contract law (Lippi et al., 2019; Tuggener et al., 2020).

Several studies (Chalkidis et al., 2018, 2019c; Hendrycks et al., 2021) explored information extraction from contracts, to extract important information such as the contracting parties, agreed payment amount, start and end dates, applicable law, etc. Other studies focus on extracting information from legislation (Cardellino et al., 2017; Angelidis et al., 2018) or court cases (Leitner et al., 2019).

Legal Question Answering (QA) is another task of interest in legal NLP, where the goal is to train models for answering legal questions (Kim et al., 2015; Ravichander et al., 2019; Kien et al., 2020; Zhong et al., 2020a,c; Louis and Spanakis, 2022). Not only is this task interesting for researchers but it could support efforts to help laypeople better understand their legal rights. In the general task setting, this requires identifying relevant legislation, case law, or other legal documents, and extracting elements of those documents that answer a particular question. A notable venue for legal QA has been the Competition on Legal Information Extraction and Entailment (COLIEE) (Kim et al., 2016; Kano et al., 2017, 2018).

More recently, there have also been efforts to pre-train Transformer-based language models on legal corpora (Chalkidis et al., 2020b; Zheng et al., 2021; Xiao et al., 2021), leading to state-of-the-art results in several legal NLP tasks, compared to models pre-trained on generic corpora.

Overall, the legal NLP literature is overwhelming, and the resources are scattered. Documentation is often not available, and evaluation measures vary across articles studying the same task. Our goal is to create the first unified benchmark to access the performance of NLP models on legal NLU. As a first step, we selected a representative group of tasks, using datasets in English that are also publicly available, adequately documented and have an appropriate size for developing modern NLP methods. We also introduce several simplifications to make the new benchmark more standardized and easily accessible, as already noted.

3 LexGLUE Tasks and Datasets

We present the Legal General Language Understanding Evaluation (LexGLUE) benchmark, a collection of datasets for evaluating model performance across a diverse set of legal NLU tasks.

3.1 Dataset Desiderata

The datasets of LexGLUE were selected to satisfy the following desiderata:

- **Language:** In this first version of LexGLUE, we only consider English datasets, which also makes experimentation easier for researchers across the globe. We hope to include other languages in future versions of LexGLUE.

- **Substance:** The datasets should check the ability of systems to understand and reason about legal text to a certain extent in order to perform tasks that are meaningful for legal practitioners.

- **Difficulty:** The performance of state-of-the-art methods on the datasets should leave large scope for improvements (cf. GLUE and SuperGLUE,

2The term "understanding" is, of course, as debatable as in NLU and GLUE, but is commonly used in NLP to refer to systems that analyze, rather than generate text.

3We reuse this term from the work of Wang et al. (2019a).
where top-ranked models now achieve average scores higher than 90%). Unlike SuperGLUE (Wang et al., 2019a), we did not rule out, but rather favored, datasets requiring domain (in our case legal) expertise.

- **Availability & Size:** We consider only publicly available datasets, documented by published articles, avoiding proprietary, untested, poorly documented datasets. We also excluded very small datasets, e.g., with fewer than 5K documents. Although large pre-trained models often perform well with relatively few task-specific training instances, newcomers may wish to experiment with simpler models that may perform disappointingly with small training sets. Small test sets may also lead to unstable and unreliable results.

### 3.2 Tasks and Datasets

LexGLUE comprises seven datasets. Table 1 shows core information for each of the LexGLUE datasets and tasks, described in detail below.¹

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Sub-domain</th>
<th>Task Type</th>
<th>Training/Dev/Test Instances</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECHR (Task A)</td>
<td>Chalkidis et al. (2019a)</td>
<td>ECHR</td>
<td>Multi-label classification</td>
<td>9,000/1,000/1,000</td>
<td>10+1</td>
</tr>
<tr>
<td>ECHR (Task B)</td>
<td>Chalkidis et al. (2021c)</td>
<td>ECHR</td>
<td>Multi-label classification</td>
<td>9,000/1,000/1,000</td>
<td>10+1</td>
</tr>
<tr>
<td>SCOTUS</td>
<td>Spaeth et al. (2020)</td>
<td>US Law</td>
<td>Multi-class classification</td>
<td>5,000/1,400/1,400</td>
<td>14</td>
</tr>
<tr>
<td>EUR-LEX</td>
<td>Chalkidis et al. (2021a)</td>
<td>EU Law</td>
<td>Multi-label classification</td>
<td>55,000/5,000/5,000</td>
<td>100</td>
</tr>
<tr>
<td>LEDGAR</td>
<td>Tuggener et al. (2020)</td>
<td>Contracts</td>
<td>Multi-class classification</td>
<td>60,000/10,000/10,000</td>
<td>100</td>
</tr>
<tr>
<td>UNFAIR-ToS</td>
<td>Lippi et al. (2019)</td>
<td>Contracts</td>
<td>Multi-label classification</td>
<td>5,532/2,275/1,607</td>
<td>8+1</td>
</tr>
<tr>
<td>CaseHOLD</td>
<td>Zheng et al. (2021)</td>
<td>US Law</td>
<td>Multiple choice QA</td>
<td>45,000/3,900/3,900</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the LexGLUE datasets, including simplifications made.

The total number of ECHR articles is currently 66. Several articles, however, cannot be violated, are rarely (or never) discussed in practice, or do not depend on the facts of a case and concern procedural technicalities. Thus, we use a simplified version of the label set (ECHR articles) in both Task A and B, including only 10 ECHR articles that can be violated and depend on the case’s facts.

**SCOTUS** The US Supreme Court (SCOTUS)⁵ is the highest federal court in the United States of America and generally hears only the most controversial or otherwise complex cases which have not been sufficiently well solved by lower courts. We release a new dataset combining information from SCOTUS opinions⁶ with the Supreme Court DataBase (SCDB)⁷ (Spaeth et al., 2020). SCDB provides metadata (e.g., decisions, issues, decision directions) for all cases (from 1946 up to 2020). We opted to use SCDB to classify the court opinions in the available 14 issue areas (e.g., Criminal Procedure, Civil Rights, Economic Activity, etc.). This is a single-label multi-class classification task (Table 1). The 14 issue areas cluster 278 issues whose focus is on the subject matter of the controversy (dispute). The SCOTUS cases are chronologically split into training (5k, 1946–1982), development (1.4k, 1982–1991), test (1.4k, 1991–2016) sets.

**EUR-LEX** European Union (EU) legislation is published in the EUR-Lex portal.⁸ All EU laws are annotated by EU’s Publications Office with multiple concepts from EuroVoc, a multilingual thesaurus maintained by the Publications Office.⁹ The current version of EuroVoc contains more than 7k concepts referring to various activities of the EU and its Member States (e.g., economics, healthcare, trade). We use the English part of the dataset of Chalkidis et al. (2021a), which comprises 65k EU laws (documents) from EUR-Lex. Given a

¹In Appendix G, we provide examples, i.e., pairs of (inputs, outputs), for all datasets and tasks.

⁵https://www.supremecourt.gov
⁶https://www.courtlistener.com
⁷http://scdb.wustl.edu
⁸http://eur-lex.europa.eu/
⁹http://eurovoc.europa.eu/
document, the task is to predict its EuroVoc labels (concepts). The dataset is chronologically split in training (55k, 1958–2010), development (5k, 2010–2012), test (5k, 2012–2016) subsets. It supports four different label granularities, comprising 21, 127, 567, 7390 EuroVoc concepts, respectively. We use the 100 most frequent concepts from level 2, which has a highly skewed label distribution and temporal concept drift (Chalkidis et al., 2021a), making it sufficiently difficult for an entry point.

**LEDGAR** Tuggener et al. (2020) introduced LEDGAR (Labeled EDGAR), a dataset for contract provision (paragraph) classification. The contract provisions come from contracts obtained from the US Securities and Exchange Commission (SEC) filings, which are publicly available from EDGAR[10] (Electronic Data Gathering, Analysis, and Retrieval system). The original dataset includes approx. 850k contract provisions labeled with 12.5k categories. Each label represents the single main topic (theme) of the corresponding contract provision, i.e., this is a single-label multi-class classification task. In LexGLUE, we use a subset of the original dataset with 80k contract provisions, considering only the 100 most frequent categories as a simplification. We split the new dataset chronologically into training (60k, 2016–2017), development (10k, 2018), and test (10k, 2019) sets.

**UNFAIR-ToS** The UNFAIR-ToS dataset (Lippi et al., 2019) contains 50 Terms of Service (ToS) from on-line platforms (e.g., YouTube, Ebay, Facebook, etc.). The dataset has been annotated on the sentence-level with 8 types of unfair contractual terms, meaning terms (sentences) that potentially violate user rights according to EU consumer law.[11] The input to a model is a sentence, the output is the set of unfair types (if any). We split the dataset chronologically into training (5.5k, 2006–2016), development (2.3k, 2017), test (1.6k, 2017) sets.

**CaseHOLD** The CaseHOLD (Case Holdings on Legal Decisions) dataset (Zheng et al., 2021) contains approx. 53k multiple choice questions about holdings of US court cases from the Harvard Law Library case law corpus. Holdings are short summaries of legal rulings that accompany referenced decisions relevant for the present case, e.g.:

> “... to act pursuant to City policy, re d 503, 506-07 (3d Cir.l985)(holding that for purposes of a class certification motion the court must accept as true all factual allegations in the complaint and may draw reasonable inferences therefrom).”

The input consists of an excerpt (or prompt) from a court decision, containing a reference to a particular case, where the holding statement (in boldface) is masked out. The model must identify the correct (masked) holding statement from a selection of five choices. We split the dataset in training (45k), development (3.9k), test (3.9k) sets, excluding samples that are shorter than 256 tokens. Chronological information is missing from CaseHOLD, thus we cannot perform a chronological re-split.

### 4 Models Considered

#### 4.1 Linear SVM

Our first baseline model is a linear Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with TF-IDF features for the top-$K$ frequent n-grams of the training set, where $n \in \{1, 2, 3\}$.

#### 4.2 Pre-trained Transformer Models

We experiment with Transformer-based (Vaswani et al., 2017) pre-trained language models, which achieve state of the art performance in most NLP tasks (Bommasani et al., 2021) and NLU benchmarks (Wang et al., 2019a). These models are pre-trained on very large unlabeled corpora to predict masked tokens (masked language modeling) and typically also to perform other pre-training tasks that still do not require any manual annotation (e.g., predicting if two sentences were adjacent in the corpus or not, dubbed next sentence prediction).

---


The pre-trained models are then fine-tuned (further trained) on task-specific (typically much smaller) annotated datasets, after adding task-specific layers. We fine-tune and evaluate the performance of the following publicly available models (Table 2).

**BERT** (Devlin et al., 2019) is the best-known pre-trained Transformer-based language model. It is pre-trained to perform masked language modeling and next sentence prediction.

**RoBERTa** (Liu et al., 2019) is also a pre-trained Transformer-based language model. Unlike BERT, RoBERTa uses dynamic masking, it eliminates the next sentence prediction pre-training task, uses a larger vocabulary, and has been pre-trained on much larger corpora. Liu et al. (2019) reported improved results on NLU benchmarks using RoBERTa, compared to BERT.

**DeBERTa** (He et al., 2021) is another improved BERT model that uses disentangled attention, i.e., four separate attention mechanisms considering the content and the relative position of each token, and an enhanced mask decoder, which explicitly considers the absolute position of the tokens. DeBERTa has been reported to outperform BERT and RoBERTa in several NLP tasks (He et al., 2021).

**Longformer** (Beltagy et al., 2020) extends Transformer-based models to support longer sequences, using sparse-attention. The latter is a combination of local (window-based) attention and global (dilated) attention that reduces the computational complexity of the model and thus can be deployed in longer documents (up to 4096 tokens). Longformer outperforms RoBERTa on long document tasks and QA benchmarks.

**BigBird** (Zaheer et al., 2020) is another sparse-attention based transformer that uses a combination of a local (window-based) attention, global (dilated), and random attention, i.e., all tokens also attend a number of random tokens on top of those in the same neighborhood (window) and the global ones. BigBird has been reported to outperform Longformer on QA and summarization tasks.

**Legal-BERT** (Chalkidis et al., 2020b) is a BERT model pre-trained on English legal corpora, consisting of legislation, contracts, and court cases. It uses the original pre-training BERT configuration. The sub-word vocabulary of Legal-BERT is built from scratch, to better support legal terminology.

**CaseLaw-BERT** (Zheng et al., 2021) is another law-specific BERT model. It also uses the original pre-training BERT configuration and has been pre-trained from scratch on the Harvard Law case corpus, which comprises 3.4M legal decisions from US federal and state courts. This model is called Custom Legal-BERT by Zheng et al. (2021). We call it CaseLaw-BERT to distinguish it from the previously published Legal-BERT of Chalkidis et al. (2020b) and to better signal that it is trained exclusively on case law (court opinions).

**Hierarchical Variants** Legal documents are usually much longer (i.e., consisting of thousands of words) than other text types (e.g., tweets, customer reviews, news articles) often considered in various NLP tasks. Thus, standard Transformer-based models that can typically process up to 512 sub-word units cannot be directly applied across all LexGLUE datasets, unless documents are severely truncated to the model’s limit. Figure 2 shows the distribution of text input length across all LexGLUE datasets. Even for Transformer-based models specifically designed to handle long text (e.g., Longformer, BigBird), handling longer legal documents remains a challenge.

Given the length of the text input in three of the seven LexGLUE tasks, i.e., ECtHR (A and B) and SCOTUS, we employ a hierarchical variant of each pre-trained Transformer-based model that has not been designed for longer text (BERT, RoBERTa, etc.)

---

12https://case.law/
DeBERTa, Legal-BERT, CaseLaw-BERT) during fine-tuning and inference. The hierarchical variants are similar to those of Chalkidis et al. (2021c). They use the corresponding pre-trained Transformer-based model to encode each paragraph of the input text independently and obtain the top-level representation $h_{[cls]}$ of each paragraph. A second-level shallow (2-layered) Transformer encoder with always the same (across BERT, RoBERTa, DeBERTa etc.) specifications (e.g., hidden units, number of attention heads) is fed with the paragraph representations to make them context-aware (aware of the surrounding paragraphs). We then max-pool over the context-aware paragraph representations to obtain a document representation, which is fed to a classification layer.\(^\text{13}\)

4.3 Task-Specific Fine-Tuning

Text Classification Tasks For EUR-LEX, LEDGAR and UNFAIR-ToS tasks, we feed each document to the pre-trained model (e.g., BERT) and obtain the top-level representation $h_{[cls]}$ of the special $[cls]$ token as the document representation, following Devlin et al. (2019). The latter goes through a dense layer of $L$ output units, one per label, followed by a sigmoid (in EUR-LEX, UNFAIR-ToS) or softmax (in LEDGAR) activation, respectively. For the two ECtHR tasks (A and B) and SCOTUS, where the hierarchical variants are employed, we feed the max-pooled (over paragraphs) document representation to a classification linear layer. The linear layer is again followed by a sigmoid (EctHR) or softmax (SCOTUS) activation.

Multiple-Choice QA Task For CaseHOLD, we convert each training (or test) instance (the prompt and the five candidate answers) into five input pairs following Zheng et al. (2021). Each pair consists of the prompt and one of the five candidate answers, separated by the special delimiter token $[sep]$. The top-level representation $h_{[cls]}$ of each pair is fed to a linear layer to obtain a logit, and the five logits are then passed through a softmax yielding a probability distribution over the five candidate answers.

4.4 Data Repository and Code

For reproducibility purposes and to facilitate future experimentation with other models, we pre-process and release all datasets on Hugging Face Datasets (Lhoest et al., 2021).\(^\text{14}\) We also release the code of our experiments, which relies on the Hugging Face Transformers (Wolf et al., 2020) library.\(^\text{16}\) Appendix A explains how to load the datasets and run experiments with our code.

5 Experiments

5.1 Experimental Set Up

For TFIDF-based linear SVM models, we use the implementation of Scikit-learn (Pedregosa et al., 2011) and grid-search for hyper parameters (number of features, $C$, and loss function). For all the pre-trained models, we use publicly available Hugging Face checkpoints.\(^\text{17}\) We use the $\ast$-base configuration of each pre-trained model, i.e., 12 Transformer blocks, 768 hidden units, and 12 attention heads. We train models with the Adam optimizer (Kingma and Ba, 2015) and an initial learning rate of $3e^{-5}$ up to 20 epochs using early stopping on development data. We use mixed precision (fp16) to decrease the memory footprint in training and gradient accumulation for all hierarchical models. The hierarchical models can read up to 64 paragraphs of 128 tokens each. We use Longformer and BigBird in default settings, i.e., Longformer uses windows of 512 tokens and a single global token ($[cls]$), while BigBird uses blocks of 64 tokens (windows: $3 \times$ block, random: $3 \times$ block, global: $2 \times$ initial block; each token attends 512 tokens in total). The batch size is 8 in all experiments. We run five repetitions with different random seeds and report the test scores based on the seed with the best scores on development data. We evaluate performance using micro-$F1$ ($\mu$-$F1$) and macro-$F1$ ($m$-$F1$) across all datasets to take into account class imbalance. For completeness, we also report the arithmetic, harmonic, and geometric mean across tasks following Shavrina and Malych (2021).\(^\text{18}\)

5.2 Experimental Results

Main Results Table 3 presents the test results for all models across all LexGLUE tasks, while Table 4

\(^{13}\)In Appendix D, we present results from preliminary experiments using the standard version of BERT for ECtHR Task A (-12.2%), Task B (-10.6%), and SCOTUS (-3.5%).

\(^{14}\)https://huggingface.co/datasets/lex_glue

\(^{15}\)https://github.com/coastalcph/lex-glue

\(^{16}\)https://huggingface.co/transformers

\(^{17}\)http://huggingface.co/models

\(^{18}\)We acknowledge that the use of scores aggregated over tasks has been criticized in general NLU benchmarks (e.g., GLUE), as models are trained with different numbers of samples, task complexity, and evaluation metrics per task. We believe that the use of a standard common metric (F1) across tasks and averaging with harmonic mean alleviate this issue.
Yang et al. (2020) proposed SMITH, a hierarchical model that considers the document structure (e.g., sentences, paragraphs) to be the main reason, i.e., in many cases full paragraphs or parts of them are not considered by the hierarchical models (limited to 64 paragraphs of 128 tokens each).

**Legal-oriented Models** Interestingly, the performance of Legal-BERT and CaseLaw-BERT, the two legal-oriented pre-trained models, is almost identical on CaseHOLD, despite the fact that CaseLaw-BERT is solely trained on US case law. On the other hand, Legal-BERT has been exposed to a wider variety of legal corpora, including EU and UK legislation, ECtHR, ECJ and US court cases, and US contracts. Legal-BERT performs as well as or better than CaseLaw-BERT on all datasets. These results suggest that domain-specific pre-training (and learning a domain-specific sub-word vocabulary) is beneficial, but over-fitting a specific (niche) sub-domain (e.g., US case law), similarly to Zheng et al. (2021), has no benefits.

### Table 3: Test results for all examined models across LexGLUE tasks. In starred datasets, we use the hierarchical variant of each model, except for Longformer and BigBird, discussed in Section 4.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>ECHR (A)*</th>
<th>ECHR (B)*</th>
<th>SCOTUS*</th>
<th>EUR-LEX</th>
<th>LEDGAR</th>
<th>UNFAIR-ToS</th>
<th>CaseHOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>µ-F₁</td>
<td>m-F₁</td>
<td>µ-F₁</td>
<td>m-F₁</td>
<td>µ-F₁</td>
<td>m-F₁</td>
<td>µ-F₁</td>
</tr>
<tr>
<td>TFIDF+SVM</td>
<td>64.5</td>
<td>51.7</td>
<td>74.6</td>
<td>65.1</td>
<td>78.2</td>
<td>69.5</td>
<td>71.3</td>
</tr>
<tr>
<td>BERT</td>
<td>71.2</td>
<td>63.6</td>
<td>79.7</td>
<td>73.4</td>
<td>68.3</td>
<td>58.3</td>
<td>71.4</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>69.2</td>
<td>59.0</td>
<td>77.3</td>
<td>68.9</td>
<td>71.6</td>
<td>62.0</td>
<td>71.9</td>
</tr>
<tr>
<td>DeBERTa</td>
<td>70.0</td>
<td>60.8</td>
<td>78.8</td>
<td>71.0</td>
<td>71.1</td>
<td>62.7</td>
<td>72.1</td>
</tr>
<tr>
<td>Longformer</td>
<td>69.9</td>
<td>64.7</td>
<td>79.4</td>
<td>71.7</td>
<td>72.9</td>
<td>64.0</td>
<td>71.6</td>
</tr>
<tr>
<td>BigBird</td>
<td>70.0</td>
<td>62.9</td>
<td>78.8</td>
<td>70.9</td>
<td>72.8</td>
<td>62.0</td>
<td>71.5</td>
</tr>
<tr>
<td>Legal-BERT</td>
<td>70.0</td>
<td>64.0</td>
<td>80.4</td>
<td>74.7</td>
<td>76.4</td>
<td>66.5</td>
<td>72.1</td>
</tr>
<tr>
<td>CaseLaw-BERT</td>
<td>69.8</td>
<td>62.9</td>
<td>78.8</td>
<td>70.3</td>
<td>76.6</td>
<td>65.9</td>
<td>70.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>A-Mean</th>
<th>H-Mean</th>
<th>G-Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>µ-F₁</td>
<td>m-F₁</td>
<td>µ-F₁</td>
</tr>
<tr>
<td>BERT</td>
<td>77.8</td>
<td>69.5</td>
<td>76.7</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>78.3</td>
<td>68.7</td>
<td>76.8</td>
</tr>
<tr>
<td>DeBERTa</td>
<td>78.3</td>
<td>69.7</td>
<td>77.4</td>
</tr>
<tr>
<td>Longformer</td>
<td>78.5</td>
<td>70.5</td>
<td>77.5</td>
</tr>
<tr>
<td>BigBird</td>
<td>78.2</td>
<td>69.6</td>
<td>77.2</td>
</tr>
<tr>
<td>Legal-BERT</td>
<td>79.8</td>
<td>72.0</td>
<td>78.9</td>
</tr>
<tr>
<td>CaseLaw-BERT</td>
<td>79.4</td>
<td>70.9</td>
<td>78.5</td>
</tr>
</tbody>
</table>

Table 3: Test scores aggregated over tasks: arithmetic (A), harmonic (H), and geometric (G) mean.

Table 4: Test scores aggregated over tasks: arithmetic (A), harmonic (H), and geometric (G) mean.

An exceptional case of the dominance of the pre-trained Transformer models is the SCOTUS dataset, where the TFIDF-based linear SVM performs best. We suspect the large size of the SCOTUS opin- ions (Figure 2) to be the main reason, i.e., in many cases full paragraphs or parts of them are not considered by the hierarchical models (limited to 64 paragraphs of 128 tokens each).

**6 Vision – Future Considerations**

Beyond the scope of this work and the examined baseline models, we identify four major factors that could potentially advance the state of the art in LexGLUE and legal NLP more generally:

**Long Documents:** Several Transformer-based models (Beltagy et al., 2020; Zaheer et al., 2020; Liu et al., 2022) have been proposed to handle long documents by exploring sparse attention mechanisms. These models can handle sequences up to 4096 sub-words, which is largely exceeded in three out of seven LexGLUE tasks (Figure 2). Contrary, the hierarchical model of Section 4.2 can handle sequences up to 8192 sub-words in our experiments, but a part of the model (the additional Transformer blocks that make the paragraph embeddings aware of the other paragraphs) is not pre-trained, which possibly negatively affects performance.

**Structured Text:** Current models for long documents, like Longformer and BigBird, do not consider the document structure (e.g., sentences, paragraphs, sections). For example, window-based attention may consider a sequence of sentences across paragraph boundaries or even consider truncated sentences. To exploit the document structure, Yang et al. (2020) proposed SMITH, a hierarchi-
cal Transformer model that hierarchically encodes increasingly larger blocks (e.g., words, sentences, documents). SMITH is very similar to the hierarchical model of Section 4.2, but it is pre-trained end-to-end with two objectives: token-level masked and sentence block language modeling.

Large-scale Legal Pre-training: Recent studies (Chalkidis et al., 2020b; Zheng et al., 2021; Bambroo and Awasthi, 2021; Xiao et al., 2021) introduced language models pre-trained on legal corpora, but of relatively small sizes, i.e., 12–36 GB. In the work of Zheng et al. (2021), the pre-training corpus covered only a narrowly defined area of legal documents, US court opinions. The same applies to Lawformer (Xiao et al., 2021), which was pre-trained on Chinese court opinions. Future work could curate and release a legal version of the C4 corpus (Raffel et al., 2020), containing multijurisdictional legislation, court decisions, contracts and legal literature at a size of hundreds of GBs. Given such a corpus, a large language model capable of processing long structured text could be pre-trained and it might excel in LexGLUE.

Even Larger Language Models: Scaling up the capacity of pre-trained models has led to increasingly better results in general NLU benchmarks (Kaplan et al., 2020), and models have been scaled up to billions of parameters (Brown et al., 2020; Raffel et al., 2020; He et al., 2021). In Appendix E, we observe that using the large version of RoBERTa leads to substantial performance improvements compared to the base version. The results are comparable or better - in some cases compared to the legal-oriented language models (Legal-BERT, CaseLaw-BERT). Considering that the two legal-oriented models are much smaller and have been pre-trained with (5–10×) less data (Section 2), we have a strong indication for performance gains by pre-training larger legal-oriented models using larger legal corpora.

7 Limitations and Future Work

Although, our benchmark inevitably cannot cover “everything in the whole wide (legal) world” (Raji et al., 2021), we include a representative collection of English datasets that also ground to a certain degree in practically interesting applications.

In its current version, LexGLUE can only be used to evaluate English models. As legal documents are typically written in the official language of the particular country of origin, there is an increasing need for developing models for other languages. The current scarcity of datasets in other languages (with the exception of Chinese) makes a multilingual extension of LexGLUE challenging, but an interesting avenue for future research.

Beyond language barriers, legal restrictions currently inhibit the creation of more datasets. Important document types, such as contracts and scholarly publications are protected by copyright or considered trade secrets. As a result, their owners are concerned with data-leakage when they are used for model training and evaluation. Providing both legal and technical solutions, e.g., using privacy-aware infrastructure and models (Downie, 2004; Feyisetan et al., 2020) is a challenge to be addressed.

Access to court decisions can also be hindered by bureaucratic inertia, outdated technology and data protection concerns, which collectively result in these otherwise public decisions not being publicly available (Pah et al., 2020). While the anonymization of personal data provides a solution to this problem, it is itself an open challenge for legal NLP (Jana and Biemann, 2021). In lack of suitable datasets and benchmarks, we have refrained from including anonymization in this version of LexGLUE, but plan to do so at a later stage.

Another limitation of the current version of LexGLUE is that human evaluation is missing. All datasets rely on ground truth labels automatically extracted from data (e.g., court decisions) produced as part of official judicial or archival procedures. These resources should be highly reliable (valid), but we cannot statistically assess their quality. In the future, re-annotating part of the datasets with multiple legal experts would provide an estimation of human level performance and inter-annotator agreement, though the cost would be high, because of the required legal expertise.

While LexGLUE offers a much needed unified testbed for legal NLU, there are several other critical aspects that need to be studied carefully. These include multi-disciplinary research to better understand the limitations and challenges of applying NLP to law (Binns, 2020), while also considering fairness and robustness (Angwin et al., 2016; Dressel and Farid, 2018; Baker Gillis, 2021; Wang et al., 2021; Chalkidis et al., 2022), and broader legal considerations of AI technologies in general (Schwemer et al., 2021; Tsarapatsanis and Aletras, 2021; Delacroix, 2022).
Acknowledgments

This work was partly funded by the Innovation Fund Denmark (IFD)\(^1\) under File No. 0175-00011A and by the German Federal Ministry of Education and Research (BMBF) kmu-innovativ program under funding code 01IS18085. We would like to thank Desmond Elliott for providing valuable feedback (baselines for truncated documents presented in Appendix D), Xiang Dai and Joel Niklaus for reviewing and pointing out issues in the new resources (code, datasets).

Ethics Statement

Original Work Attribution

All datasets included in LexGLUE, except SCOTUS, are publicly available and have been previously published. If datasets or the papers that introduced them were not compiled or written by ourselves, we referenced the original work and encourage LexGLUE users to do so as well. In fact, we believe this work should only be referenced, in addition to citing the original work, when experimenting with multiple LexGLUE datasets and using the LexGLUE evaluation infrastructure. Otherwise only the original work should be cited.

Social Impact

We believe that this work does not contain any grounds for ethical concerns. A transparent and rigorous benchmark for NLP in the legal domain might serve as an orientation for scholars and industry researchers. As a result, the capabilities of tools that are trained using natural language data from the legal domain will become clearer, thereby helping their users to better understand them. This increased certainty would also raise the awareness within research and industry communities to potential risks associated with the use of these tools. We regard this contribution to a more realistic, more informed discussion as an important use case of the work presented. Ideally, it could help both beginners and seasoned professionals to understand the limitations of using NLP tools in the legal domain and thereby prevent exaggerated expectations and potential applications that might risk endangering fundamental rights or the rule of law. We currently cannot imagine use cases of this particular work that would lead to ethical concerns or potential harm (Tsarapatsanis and Aletras, 2021).

LexGLUE comprises seven datasets: ECtHR Task A and B, SCOTUS, EUR-LEX, LEDGAR, UNFAIR-ToS, and CaseHOLD that are available for re-use and re-share with appropriate attribution. The data is in general partially anonymized in accordance with the applicable national law. The data is considered to be in the public sphere from a privacy perspective. This is a very sensitive matter, as the courts try to keep a balance between transparency (the public’s right to know) and privacy (respect for private and family life).

ECtHR contains personal data of the parties and other people involved in the legal proceedings. Its data is processed and made public in accordance with the European data protection laws. This includes either implied consent or legitimate interest to process the data for research purposes. As a result, their processing by us or other future users of the benchmark is not likely to raise ethical concerns.

SCOTUS contains personal data of a similar nature. Again, the data is processed and made available by the US Supreme Court, whose proceedings are public. While this ensures compliance with US law, it is very likely that similarly to the ECtHR any processing could be justified by either implied consent or legitimate interest under European law.

EUR-LEX by contrast is merely a collection of legislation material and therefore not likely to contain personal data, except signatory information (e.g., president of EC). It is openly published by the European Union and processed by the EU’s Publication Office. In addition, since our work qualifies as research, it is privileged pursuant to Art. 6 (1) (f) GDPR.

LEDGAR contains publicly available contract provisions published in the EDGAR database of the US Securities and Exchange Commission (SEC). As far as personal information might be contained, it should equally fall into the public sphere and be covered by research privilege. Our processing does not focus on personal information at all, rather attributing content labels to provisions.

UNFAIR-ToS contains Terms of Services from business entities such as YouTube, Ebay, Facebook, etc., which makes it unlikely for the data to include personal information. These companies keep user data separate from contractual provisions, so to the best of our knowledge not contained in this dataset.

CaseHOLD contains parts of legal decisions.

\(^1\)https://innovationsfonden.dk/en
from US Court decisions, obtained from the Harvard library case law corpus. All of the decisions were previously published in compliance with US law. In addition, most instances (case snippets) are too short to contain identifiable information. Should such data be contained, their processing would equally be covered either by implicit consent or a public interest exception. We use all datasets in accordance with copyright terms and under the licenses set forth by their creators.

Limitations & Potential Harms

We have not employed any crowd-workers or annotators for this work. The paper outlines the main limitations with regard to speaker population (English) and generalizability in a dedicated section (Section 7). As a benchmark paper, our claims naturally match the results of the experiments, which – given the current detail of instructions – should be easily reproduced. We provide several ways of accessing the datasets and running the experiments both with and without Hugging Face infrastructure.

We do not currently foresee any potential harms for vulnerable or marginalized populations and we do not use, to the best of our knowledge, any identifying characteristics for populations of these kinds.

References


Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks. ProPublica.


Association for Computational Linguistics, Dublin, Ireland.


Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).


A Datasets, Code, and Participation

Where are the datasets? We provide access to LexGLUE on Hugging Face Datasets (Lhoest et al., 2021) at https://huggingface.co/datasets/lex_glue. For example, to load the SCOTUS dataset, you first simply install the datasets Python library and then make the following call:

```python
from datasets import load_dataset
dataset = load_dataset('lex_glue', task='scotus')
```

How do I run experiments? To make reproducible the results of the already examined models or future models even easier, we release our code on GitHub (https://github.com/coastalrph/lex-glue). In that repository (in the folder/experiments), there are Python scripts, relying on the Hugging Face Transformers library (Wolf et al., 2020), to run and evaluate any Transformer-based model (e.g., BERT, RoBERTa, LegalBERT, and their hierarchical variants, as well as, Longformer, and BigBird). We also provide bash scripts to replicate the experiments for each dataset with 5 random seeds, as we did for the reported results.

B No labeling as an additional class

In ECtHR Tasks A & B and UNFAIR-ToS, there are unlabeled samples. Concretely, in ECtHR Task A, a possible event is no violation, i.e., the court ruled that the defendant did not violate any ECHR article. Contrary, no violation is not a possible event in the original ECtHR Task B dataset, i.e., at least a single ECHR article is allegedly violated (considered by the court) in every case; however, there is such a rare scenario after the simplifications we introduced, i.e., some cases were originally labeled only with rare labels that were excluded from our benchmark (Section 3.2). In UNFAIR-ToS, the vast majority of sentences are not labeled with any type of unfairness (unfair term against users), i.e., most sentences do not raise any questions of possible violations of the European consumer law.

In multi-label classification, the set of labels per instance is represented as a one-hot vector $Y = [y_1, y_2, \ldots, y_L]$, where $y_i = 1$ if the instance is labeled with the $i$-th class, and $y_i = 0$ otherwise. If an instance is not labeled with any class, its $Y$ includes only zeros. During training, binary cross-entropy correctly penalizes such instances, if the predictions ($\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L]$) diverge from zeros. During evaluation, however, the $F_1$-score ($F_1 = \frac{TP}{TP+FP+FN}$) ignores instances with $Y = \hat{Y} = [0, 0, \ldots, 0]$, because it considers only the true positives (TP), false positives (FP), and false negatives (FN), and instances where $Y = \hat{Y} = [0, 0, \ldots, 0]$ contribute no TPs, FPs, FNs. In order to make $F_1$ sensitive to the correct labeling of such examples, during evaluation (not training) we include an additional label ($y_0$ or $\hat{y}_0$) in both targets ($Y$) and predictions ($\hat{Y}$), whose value is 1 (positive) if the instance (without $y_0$, $\hat{y}_0$) $Y$ and $\hat{Y}$ are $Y = [0, 0, \ldots, 0]$ or $\hat{Y} = [0, 0, \ldots, 0]$, respectively, and 0 (negative) otherwise. This is particularly important for proper evaluation, as across three datasets a considerable portion of the examples are unlabeled (11.5% in ECtHR Task A, 1.6% in ECtHR Task B, and 95.5% in UNFAIR-ToS).

C Additional Results

Tables 5 and 6 show development results for all examined models across datasets. We report the mean and standard deviations ($\pm$) for the three seeds (among the five used) with the best development scores per model to exclude catastrophic failures, i.e., runs with severely low performance. The standard deviations are relatively low across models and datasets (up to 0.5% for $\mu$-F1 and up to 1% for m-F1). The development results are generally higher compared to the test ones (cf. Table 3) in many cases, as one would expect.

Table 7 reports training times per dataset and model; both the time per epoch ($T/e$), and the total training time ($T$) across all epochs. All full-attention BERT models, except Longformer and

References...


Big-Bird, have comparable times with the exception of DeBERTa that has four separate attention mechanisms. We observe that when the hierarchical variant of these models is deployed, i.e., in ECtHR tasks and SCOTUS, it is approximately twice (2×) as fast compared to Longformer and BigBird.

### D Use of 512-token BERT models

In Figure 3, we show results for the standard BERT model of Devlin et al. (2019), which can process up to 512 tokens, compared to its hierarchical variant (Section 4.2), which can process up to 64×128 tokens. We observe that across all datasets that contain long documents (ECtHR A & B, SCOTUS, cf. Fig. 2(a)), the hierarchical variant clearly outperforms the standard model fed with truncated documents (ECtHR A: +10.2% p.p., ECtHR B: 7.5% p.p., SCOTUS: 4.9% p.p.). Compared to the ECtHR tasks, the gains are lower in SCOTUS, a topic classification task where long-range reasoning is not needed; by contrast, for ECtHR multiple distant facts need to be combined. Based on these results, we conclude that using severely truncated documents is not a plausible option for LexGLUE, and other directions for processing long documents should be considered in the future, ideally fully pre-trained hierarchical models, contrary to our semi-pre-trained hierarchical models (Section 6).

### E Use of Roberta Large

We additionally evaluate RoBERTa-large, i.e., 24 Transformer blocks, 1024 hidden units, and 18 attention heads, to better understand the dynamics between domain specificity and model size. In this

---

### Table 5: Development $\mu$-F$_1$ results for all examined models across all LexGLUE tasks. We report the mean and standard deviations (±) for the three seeds with the best development scores per model. In starred datasets, we use the hierarchical variant of each model, except for Longformer and BigBird, as discussed in Section 4.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>ECtHR (A)*</th>
<th>ECtHR (B)*</th>
<th>SCOTUS*</th>
<th>EUR-LEX</th>
<th>LEDGAR</th>
<th>UNFAIR-ToS</th>
<th>CaseHOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>71.0 ± 0.7</td>
<td>79.6 ± 0.5</td>
<td>72.7 ± 0.2</td>
<td>77.3 ± 0.2</td>
<td>87.9 ± 0.1</td>
<td>95.5 ± 0.0</td>
<td>72.8 ± 0.1</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>70.4 ± 0.5</td>
<td>78.4 ± 0.7</td>
<td>76.9 ± 0.6</td>
<td>77.6 ± 0.0</td>
<td>88.1 ± 0.1</td>
<td>94.8 ± 0.2</td>
<td>74.1 ± 0.2</td>
</tr>
<tr>
<td>DeBERTa</td>
<td>69.3 ± 0.7</td>
<td>79.0 ± 0.3</td>
<td>76.1 ± 0.5</td>
<td>78.2 ± 0.1</td>
<td>88.3 ± 0.2</td>
<td>95.5 ± 0.1</td>
<td>73.8 ± 0.1</td>
</tr>
<tr>
<td>Longformer</td>
<td>71.0 ± 0.3</td>
<td>80.4 ± 0.9</td>
<td>76.9 ± 0.0</td>
<td>77.5 ± 0.0</td>
<td>88.1 ± 0.2</td>
<td>95.1 ± 0.2</td>
<td>73.9 ± 0.2</td>
</tr>
<tr>
<td>BigBird</td>
<td>71.0 ± 0.2</td>
<td>80.1 ± 0.5</td>
<td>75.9 ± 0.2</td>
<td>77.3 ± 0.1</td>
<td>88.0 ± 0.1</td>
<td>95.2 ± 0.4</td>
<td>73.7 ± 0.2</td>
</tr>
<tr>
<td>Legal-BERT</td>
<td>71.9 ± 0.4</td>
<td>79.8 ± 0.2</td>
<td>80.4 ± 0.3</td>
<td>77.6 ± 0.1</td>
<td>88.5 ± 0.0</td>
<td>95.1 ± 0.2</td>
<td>76.4 ± 0.3</td>
</tr>
<tr>
<td>CaseLaw-BERT</td>
<td>72.1 ± 0.3</td>
<td>79.6 ± 0.0</td>
<td>81.3 ± 0.6</td>
<td>77.2 ± 0.1</td>
<td>88.4 ± 0.2</td>
<td>95.3 ± 0.4</td>
<td>77.4 ± 0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>ECtHR (A)*</th>
<th>ECtHR (B)*</th>
<th>SCOTUS*</th>
<th>EUR-LEX</th>
<th>LEDGAR</th>
<th>UNFAIR-ToS</th>
<th>CaseHOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>65.4 ± 1.2</td>
<td>74.8 ± 0.6</td>
<td>65.9 ± 0.8</td>
<td>62.6 ± 0.8</td>
<td>81.8 ± 0.1</td>
<td>75.8 ± 1.3</td>
<td>72.8 ± 0.1</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>65.4 ± 0.2</td>
<td>74.2 ± 1.1</td>
<td>69.5 ± 0.8</td>
<td>63.5 ± 0.4</td>
<td>81.9 ± 0.2</td>
<td>74.4 ± 0.7</td>
<td>74.1 ± 0.2</td>
</tr>
<tr>
<td>DeBERTa</td>
<td>63.5 ± 0.9</td>
<td>74.0 ± 0.4</td>
<td>68.4 ± 0.8</td>
<td>63.6 ± 0.3</td>
<td>82.0 ± 0.5</td>
<td>77.1 ± 1.2</td>
<td>73.8 ± 0.1</td>
</tr>
<tr>
<td>Longformer</td>
<td>65.5 ± 1.6</td>
<td>77.7 ± 1.0</td>
<td>70.4 ± 0.5</td>
<td>63.8 ± 0.5</td>
<td>82.0 ± 0.3</td>
<td>75.2 ± 1.2</td>
<td>73.9 ± 0.2</td>
</tr>
<tr>
<td>BigBird</td>
<td>65.8 ± 1.1</td>
<td>74.1 ± 0.5</td>
<td>69.1 ± 0.2</td>
<td>63.0 ± 0.3</td>
<td>81.7 ± 0.2</td>
<td>76.5 ± 1.8</td>
<td>73.7 ± 0.2</td>
</tr>
<tr>
<td>Legal-BERT</td>
<td>68.0 ± 0.2</td>
<td>76.1 ± 0.5</td>
<td>72.7 ± 0.2</td>
<td>62.0 ± 0.9</td>
<td>82.2 ± 0.3</td>
<td>76.9 ± 1.3</td>
<td>76.4 ± 0.3</td>
</tr>
<tr>
<td>CaseLaw-BERT</td>
<td>67.1 ± 0.7</td>
<td>74.6 ± 0.5</td>
<td>74.0 ± 1.2</td>
<td>62.9 ± 0.3</td>
<td>82.3 ± 0.3</td>
<td>76.5 ± 0.3</td>
<td>77.4 ± 0.2</td>
</tr>
</tbody>
</table>

---

**Figure 3:** Development $m$-F$_1$ scores of standard BERT (up to 512 tokens) and its hierarchical variant (Section 4.2, 64×128 tokens) in ECtHR (Task A, B) and SCOTUS, i.e., the datasets with long documents. Light blue denotes the average score across 5 runs for the hierarchical variant (used in Table 3 for these datasets), while dark blue corresponds to standard BERT (not used in Table 3 for these datasets). The error bars show the standard error.
Table 7: Training time in total (T) and per epoch (T/e) across LexGLUE tasks. In starred datasets, we use the hierarchical variant of each model, except for Longformer and BigBird, as described in Section 4.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>ECtHR (A)*</th>
<th>ECtHR (B)*</th>
<th>SCOTUS*</th>
<th>EUR-LEX</th>
<th>LEDGAR</th>
<th>CaseHOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T T/e</td>
<td>T T/e</td>
<td>T T/e</td>
<td>T T/e</td>
<td>T T/e</td>
<td>T T/e</td>
</tr>
<tr>
<td>BERT</td>
<td>3h 42m 28m</td>
<td>3h 9m 28m</td>
<td>1h 24m 11m</td>
<td>3h 36m 19m</td>
<td>6h 9m 21m</td>
<td>4h 24m 24m</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>4h 11m 27m</td>
<td>3h 43m 27m</td>
<td>2h 46m 17m</td>
<td>3h 36m 19m</td>
<td>6h 22m 21m</td>
<td>4h 21m 24m</td>
</tr>
<tr>
<td>DeBERTa</td>
<td>7h 43m 46m</td>
<td>6h 48m 46m</td>
<td>3h 42m 29m</td>
<td>5h 34m 36m</td>
<td>9h 29m 40m</td>
<td>6h 42m 45m</td>
</tr>
<tr>
<td>Longformer</td>
<td>6h 47m 56m</td>
<td>7h 31m 56m</td>
<td>6h 27m 34m</td>
<td>11h 10m 45m</td>
<td>15h 47m 50m</td>
<td>4h 45m 30m</td>
</tr>
<tr>
<td>BigBird</td>
<td>8h 41m 1h 2m</td>
<td>8h 17m 1h 2m</td>
<td>5h 51m 37m</td>
<td>3h 57m 24m</td>
<td>8h 13m 27m</td>
<td>6h 4m 49m</td>
</tr>
<tr>
<td>Legal-BERT</td>
<td>3h 52m 28m</td>
<td>3h 2m 28m</td>
<td>2h 2m 17m</td>
<td>3h 22m 19m</td>
<td>5h 23m 21m</td>
<td>4h 13m 23m</td>
</tr>
<tr>
<td>CaseLaw-BERT</td>
<td>3h 2m 28m</td>
<td>2h 57m 28m</td>
<td>2h 34m 34m</td>
<td>3h 40m 19m</td>
<td>6h 8m 21m</td>
<td>4h 21m 24m</td>
</tr>
</tbody>
</table>

Table 8: Development and test results across LexGLUE tasks. In starred datasets, we use the hierarchical variant of each model, discussed in Section 4.2. (B) and (L) denote the base and large version of RoBERTa, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>ECtHR (A)*</th>
<th>ECtHR (B)*</th>
<th>SCOTUS*</th>
<th>EUR-LEX</th>
<th>LEDGAR</th>
<th>UNFAIR-ToS</th>
<th>CaseHOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>μ-F1 m-F1</td>
<td>μ-F1 m-F1</td>
<td>μ-F1 m-F1</td>
<td>μ-F1 m-F1</td>
<td>μ-F1 m-F1</td>
<td>μ-F1 m-F1</td>
<td>μ-F1 m-F1</td>
</tr>
<tr>
<td>Results on Development Set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa (B)</td>
<td>70.6 65.7</td>
<td>79.3 75.8</td>
<td>77.5 64.1</td>
<td>77.6 70.4</td>
<td>88.0 82.1</td>
<td>94.6 75.2</td>
<td>74.3</td>
</tr>
<tr>
<td>RoBERTa (L)</td>
<td>72.7 69.3</td>
<td>81.1 77.0</td>
<td>74.6 56.9</td>
<td>80.7 74.5</td>
<td>88.5 82.8</td>
<td>95.8 80.3</td>
<td>76.8</td>
</tr>
<tr>
<td>Legal-BERT</td>
<td>72.5 68.2</td>
<td>79.7 76.8</td>
<td>77.6 63.3</td>
<td>80.8 72.9</td>
<td>88.5 82.6</td>
<td>95.3 78.2</td>
<td>76.6</td>
</tr>
<tr>
<td>CaseLaw-BERT</td>
<td>71.8 67.7</td>
<td>79.5 74.9</td>
<td>77.3 63.1</td>
<td>82.1 75.6</td>
<td>88.7 82.7</td>
<td>95.7 76.9</td>
<td>77.7</td>
</tr>
</tbody>
</table>

| Results on Test Set |
| RoBERTa (B)   | 69.2 59.0  | 77.3 68.9  | 71.6 62.0  | 71.9 57.9  | 87.9 82.3 | 95.2 79.2  | 71.4     |
| RoBERTa (L)   | 73.8 67.6  | 79.8 71.6  | 75.5 66.3  | 67.9 50.3  | 88.6 83.6 | 95.8 81.6  | 74.4     |
| Legal-BERT    | 70.0 64.0  | 80.4 74.7  | 76.4 66.5  | 72.1 57.4  | 88.2 83.0 | 96.0 83.0  | 75.3     |
| CaseLaw-BERT  | 69.8 62.9  | 78.8 70.3  | 76.6 65.9  | 70.7 56.6  | 88.3 83.0 | 96.0 82.3  | 75.4     |

In case, we use the AdamW optimizer with a 1e-5 maximum learning rate, warm-up ratio of 0.1, and a weight decay rate of 0.06, and we use a similar mini-batch size of 8 examples.20

Table 8 reports the development and test results using the seed (run) with the best development scores. We observe that using the large version of RoBERTa, dubbed RoBERTa (L), with more than 2x parameters (355M), leads to substantial performance improvements compared to the base version of RoBERTa, dubbed RoBERTa (B), in many tasks. The results are comparable, or better in some cases, compared to the legal-oriented language models (Legal-BERT, CaseLaw-BERT). Considering that the two legal-oriented models are much smaller and have been pre-trained with (5–10x) less data (Section 2), we have a strong indication for performance gains by pre-training larger legal-oriented models using larger legal corpora (Section 6).

F Other Tasks and Datasets Considered

We considered including the Contract Understanding Atticus Dataset (CUAD) (Hendrycks et al., 2021), an expertly curated dataset that comprises 510 contracts annotated with 41 valuable contractual insights (e.g., agreement date, parties, governing law). The task is formulated as a SQUAD-like question answering task, where given a question (the name of an insight) and a paragraph from the contract, the model has to identify the answer span in the paragraph.21 The original dataset follows the SQUAD v2.0 setting, including unanswerable questions. Following SQUAD v1.1 (Rajpurkar et al., 2016), we simplified the task by removing all unanswerable pairs (question, paragraph), which are the majority in the original dataset. We also excluded pairs whose answers exceeded 128 full words to alleviate the imbalance between short and long answers. We then re-split the dataset chronologically into training (5.2k, 1994–2019), development (572, 2019–2020), and test (604, 2020) sets.

Following Devlin et al. (2019), and similarly to Hendrycks et al. (2021), for each training (or test)

---

20Large models tend to be very sensitive to parameter updates, especially in the initial training steps; hence a smaller learning rate and warm up steps are very crucial.

21The question mostly resembles a prompt, rather than a natural question, as there is a closed set of 41 alternatives.
instance, we consider pairs that consist of a question and a paragraph, separated by the special delimiter token [sep]. The top-level representations $[h_1, \ldots, h_N]$, of the tokens of the paragraph are fed into a linear layer to obtain two logits per token (for the token being the start or end of the answer span), which are then passed through a softmax activation (separately for start and end) to obtain probability distributions. The tokens with the highest start and end probabilities are selected as boundaries of the answer span. We evaluated performance with token-level F1 score, similarly to SQUAD.

We trained all the models of Table 2, which scored approx. 10-20% in token-level F1, with Legal-BERT performing slightly better than the rest (+5% F1).\(^22\) In the paper that introduced CUAD (Hendrycks et al., 2021), several other measures (Precision@N% Recall, AUPR, Jaccard similarity) are used to more leniently estimate a model’s ability to approximately locate answers in context paragraphs. Through careful manual inspection of the dataset, we noticed the following points that seem to require more careful consideration.

- **Contractual insights (categories, shown in italics below) include both entity-level (short) answers (e.g., “SERVICE AGREEMENT” for Document Name, and “Imprimis Pharmaceuticals, Inc.” for Parties) and paragraph-level (long) answers (e.g., “If any of the conditions specified in Section 8 shall not have been fulfilled when and as required by this Agreement, or by the Closing Date, or waived in writing by Capital Resources, this Agreement and all of Capital Resources obligations hereunder may be canceled [...] except as otherwise provided in Sections 2, 7, 9 and 10 hereof.” for Termination for Convenience).** These two different types of answers (short and paragraph-long) seem to require different models and different evaluation measures, unlike how they are treated in the original CUAD paper.

- **Some contractual insights (categories), e.g., Parties, have been annotated with both short (e.g., “Imprimis Pharmaceuticals, Inc.”) and long (e.g., “together, Blackwell and Munksgaard shall be referred to as ‘the Publishers’.”) answers. Annotations of this kind introduce noise during both training and evaluation. For example, it becomes unclear when a short (finer/strict) or a long (loose) annotation should be taken to be the correct one.**

- **Annotations may include indirect mentions, e.g., ‘Franchisee’, ‘Service Provider’ for Parties, instead of the actual entities (the company name).**

- **Annotations may include semi-redacted text (e.g., “______, 1996” for Agreement Date), or even fully redacted text (e.g., “________” for Parties). This practice may be necessary to hide sensitive information, but for the purposes of a benchmark dataset such cases could have been excluded.**

The points above, which seem to require revisiting the annotations of CUAD, and the very low F1 scores of all models led us to exclude CUAD from LexGLUE. We also note that there is related work covering similar topics, such as Contract Element Extraction (Chalkidis and Androutsopoulos, 2017), Contractual Obligation Extraction (Chalkidis et al., 2018), and Contractual Provision Classification (Tuggener et al., 2020), where models perform much better (in terms of accuracy), relying on simpler (separate) more carefully designed tasks and much bigger datasets. Thus we believe that the points mentioned above, which blur the task definition of CUAD and introduce noise, and the limited (compared to larger datasets) number of annotations strongly affect the performance of the models on CUAD, understimating their true potential.

We also initially considered some very interesting legal Information Retrieval (IR) datasets (Locke and Zuccon, 2018; Chalkidis et al., 2021b) that aim to examine crucial real-life tasks (relevant case law retrieval, regulatory compliance). However, we decided to exclude them from the first version of LexGLUE, because they rely on processing multiple long documents and require more task-specific neural network architectures (e.g., siamese networks), and different evaluation measures. Hence, they would make LexGLUE more complex and a less attractive entry point for newcomers to legal NLP. We plan, however, to include more demanding tasks in future LexGLUE versions, as the legal NLP community will be growing.

G Dataset Examples

In Table 9, we present training examples, i.e., pairs of input(s), output(s), for LexGLUE datasets and tasks. More examples can be inspected using the dataset preview functionality provided in the online dataset card of Hugging Face.\(^23\)

\(^{22}\)F1 is one of the two official SQUAD measures. In the second one, Exact Answer Accuracy, all models scored 0%.

\(^{23}\)https://huggingface.co/datasets/lex_glue
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Input(s)</th>
<th>Output(s) / Label(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECtHR</td>
<td>Text: 12. In 1987 the applicant association published a book entitled Euskadi at war. There were four versions – Basque, English, Spanish and French – and the book was distributed in numerous countries, including France and Spain. According to the applicant association, this was a collective work containing contributions from a number of academics with specialist knowledge of the Basque Country and giving an account of the historical, cultural, linguistic and socio-political aspects of the Basque cause. It ended with a political article entitled “Euskadi at war, a promise of peace” by the Basque national liberation movement. (Choice A) The book was published in the second quarter of 1987. On 29 April 1988 a ministerial order was issued by the French Ministry of the Interior under section 14 of the Law of 29 July 1881, as amended by the decree of 6 May 1939, banning the circulation, distribution and sale of the book in France in any of its four versions on the ground that “the circulation in France of this book, which promotes separatism and vindicates recourse to violence, is likely to constitute a threat to public order”. On 6 May 1988, pursuant to the aforementioned order, the department director of the airport and border police refused to allow over two thousand copies of the book to be brought into France.</td>
<td>3 (Right to a fair trial) 6 (Freedom of expression)</td>
</tr>
<tr>
<td>SCOTUS</td>
<td>Text: 329 U.S. 29 67 S.Ct. 1 91 L.Ed. 22 CHAMPLIN REFINING CO v. UNITED STATES et al. No. 21. Argued Oct. 18, 21, 1946. Decided Nov. 18, 1946. Appeal from the District Court of the United States for the Western District of Oklahoma. Messrs. Dan Moody, of Austin, Tex., and Harry O. Glasser, of Enid, Okla., for appellant. Mr. Edward Dumbauld, of Washington, D.C., for appellees. Mr. Justice JACKSON delivered the opinion of the Court. I The Interstate Commerce Commission, acting under § 19a of the Interstate Commerce Act, ordered the appellant to furnish certain inventories, schedules, maps and charts of its pipe line property. 2 Champlin’s objections that the Act does not authorize the order, or if it be construed to do so is unconstitutional, were overruled by the Commission and again by the District Court which dismissed the company’s suit for an injunction. These questions of law are brought here by appeal.</td>
<td>7 (Economic Activity)</td>
</tr>
<tr>
<td>EUR-LEX</td>
<td>Text: Commission Regulation (EC) No 1156/2001 of 13 June 2001 fixing the export refunds on white sugar and raw sugar exported in its unaltered state THE COMMISSION OF THE EUROPEAN COMMUNITIES Having regard to the Treaty establishing the European Community, Having regard to Council Regulation (EC) No 2038/1999 of 13 September 1999 on the common organisation of the markets in the sugar sector(s), as amended by Commission Regulation (EC) No 1527/2000(s), and in particular point (a) of the second subparagraph of Article 18(5) thereof, Whereas: (1) Article 18 of Regulation (EC) No 2038/1999 provides that the difference between quotations or prices on the world market for the products listed in Article 1(1)(a) of that Regulation and prices for those products within the Community may be covered by an export refund. (2) Regulation (EC) No 2038/1999 provides that when refunds on white and raw sugar, undenatured and exported in its unaltered state, are being fixed account must be taken of the situation on the Community and world markets in sugar and in particular of the price and cost factors (…)</td>
<td>28 (Trade Policy), 93 (Beverages and Sugar), 94 (Foodstuff)</td>
</tr>
<tr>
<td>LEDGAR</td>
<td>Text: The validity or unenforceability of any provision or provisions of this Agreement shall not affect the validity or enforceability of any other provision hereof, which will remain in full force and effect. Should a court or other body of competent jurisdiction determine that any provision of this Agreement is excessive in scope or otherwise illegal, invalid, void or unenforceable, such provision shall be adjusted rather than voided, if possible, so that it is enforceable to the maximum extent possible. (Contract by Using) 79 (Severability)</td>
<td></td>
</tr>
<tr>
<td>UNFAIR-ToS</td>
<td>Text: By creating a tinder account or by using the tinder imessage app (“tinder stacks”), whether through a mobile device, mobile application or computer (collectively, the “service”) you agree to be bound by (i) these terms of use, (ii) our privacy policy and safety tips, each of which is incorporated by reference into this agreement, and (ii) any terms disclosed and agreed to by you if you purchase additional features, products or services we offer on the service (collectively, this “agreement”).</td>
<td>4 (Contract by Using)</td>
</tr>
<tr>
<td>CaseHOLD</td>
<td>Case: Deapen’s cohorts, the cohort would be a “victim” of making the bomb. Further, firebombs are inherently dangerous. There is no peaceful purpose for making a bomb. Felony offenses that involve explosives qualify as “violent crimes” for purposes of enhancing the sentences of career offenders. See 18 U.S.C. § 924(e)(2)(B)(ii) (defining a “violent felony” as “any crime punishable by imprisonment for a term exceeding one year... that involves use of explosives”). Courts have found possession of a bomb to be a crime of violence based on the lack of a nonviolent purpose for a bomb and the fact that, by its very nature, there is a substantial risk that the bomb would be used against the person or property of another. See United States v. Newman, 125 F.3d 863 (10th Cir.1997) (unpublished) (HOLDING); United States v. Dodge, 846 F.Supp. 181 (Economic Activity) 0 (Choice A)</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Training examples (pairs of inputs, outputs) for LeXGLUE datasets and tasks.