LexGLUE: A Benchmark Dataset for Legal Language Understanding in English

Ilias Chalkidis [α*] **Abhik Jana** [β] **Dirk Hartung** [γ δ] **Michael Bommarito** [δ]
Ion Androutsopoulos [ϵ] **Daniel Martin Katz** [γ δ ζ] **Nikolaos Aletras** [η]
[α] University of Copenhagen, Denmark [β] Universität Hamburg, Germany
[γ] Bucerius Law School, Hamburg, Germany [δ] CodeX, Stanford Law School, United States
[ϵ] Athens University of Economics and Business, Greece [η] University of Sheffield, UK
[ζ] Illinois Tech – Chicago Kent College of Law, United States

Abstract

Law, interpretations of law, legal arguments, agreements, etc. 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 (Nallapati 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).

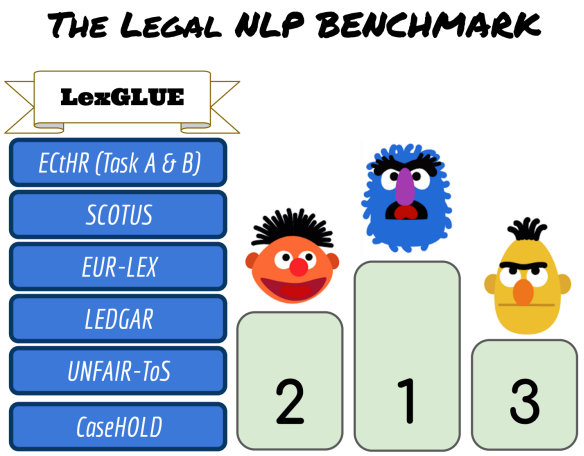

Figure 1: LexGLUE: A new benchmark dataset to evaluate the capabilities of NLU models on legal text.

* Corresponding author: ilias.chalkidis@di.ku.dk

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 senses 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).

We anticipate that more datasets, tasks, and languages will be added in later versions of LexGLUE.[1] As more legal NLP datasets become available, we also plan to favor datasets checked thoroughly for validity (scores reflecting real-life performance), annotation quality, statistical power, and social bias (Bowman and Dahl, 2021).

As in GLUE and SuperGLUE (Wang et al., 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, experiment with and evaluate their performance (Section 4.5).

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

[1] See https://nllpw.org/resources/ and https://github.com/thunlp/LegalPapers for lists of papers, datasets, and other resources related to NLP for legal text.

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 task has 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). 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

3.1 Benchmark Characteristics & Desiderata

We present the Legal General Language Understanding[2] Evaluation (LexGLUE) benchmark, a collection of datasets for evaluating model performance across a diverse set of legal NLU tasks. LexGLUE has the following main characteristics:

- **Language:** We only consider English datasets, to make experimentation easier for researchers across the globe.
- **Substance:**[3] 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, 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.

[2]The 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.

[3]We reuse this term from the work of Wang et al. (2019a).

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

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

- **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

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

ECtHR Tasks A & B The European Court of Human Rights (ECtHR) hears allegations that a state has breached human rights provisions of the European Convention of Human Rights (ECHR). We use the dataset of Chalkidis et al. (2019a, 2021c), which contains approx. 11K cases from the ECtHR public database. The cases are chronologically split into training (9k, 2001–2016), development (1k, 2016–2017), and test (1k, 2017–2019). For each case, the dataset provides a list of *factual* paragraphs (facts) from the case description. Each case is mapped to *articles* of the ECHR that were violated (if any). In Task A, the input to a model is the list of facts of a case, and the output is the set of violated articles. In the most recent version of the dataset (Chalkidis et al., 2021c), each case is also mapped to articles of ECHR that were *allegedly* violated (considered by the court). In Task B, the input is again the list of facts of a case, but the output is the set of allegedly violated articles.

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)[4] 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 combine information from SCOTUS opinions[5] with the Supreme Court DataBase (SCDB)[6] (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 EUR-Lex portal.[7] All EU laws are annotated by EU's Publications Office with multiple concepts from the EuroVoc thesaurus, a multilingual thesaurus maintained by the Publications Office.[8] The current version of EuroVoc contains more than 7k concepts referring to various activities of the EU and its Member States (e.g., economics, health-care, 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 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

[4] https://www.supremecourt.gov
[5] https://www.courtlistener.com
[6] http://scdb.wustl.edu
[7] http://eur-lex.europa.eu/
[8] http://eurovoc.europa.eu/

Method	Source	# Params	Vocab. Size	Max Length	Pretrain Specs	Pre-training Corpora	
BERT	(Devlin et al., 2019)	110M	32K	512	1M / 256	(16GB)	Wiki, BC
RoBERTa	(Liu et al., 2019)	125M	50K	512	100K / 8K	(160GB)	Wiki, BC, CC-News, OWT
DeBERTa	(He et al., 2021)	139M	50K	512	1M / 256	(160GB)	Wiki, BC, CC-News, OWT
Longformer*	(Beltagy et al., 2020)	149M	50K	4096	65K / 64	(160GB)	Wiki, BC, CC-News, OWT
BigBird*	(Zaheer et al., 2020)	127M	50K	4096	1M / 256	(160GB)	Wiki, BC, CC-News, OWT
Legal-BERT	(Chalkidis et al., 2020b)	110M	32K	512	1M /256	(12GB)	Legislation, Court Cases, Contracts
CaseLaw-BERT	(Zheng et al., 2021)	110M	32K	512	2M /256	(37GB)	US Court Cases

Table 2: Key specifications of the examined models. We report the number of parameters, the size of vocabulary, the maximum sequence length, the core pre-training specifications (training steps and batch size) and the training corpora (cf. OWT = OpenWebText, BC = BookCorpus). * Models have been warmed started from RoBERTa.

use the 100 most frequent concepts from level 2, which has a highly skewed label distribution and temporal concept drift, making it sufficiently difficult for an entry point baseline.

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[9] (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* (sentences), meaning terms that potentially violate user rights according to the European consumer law.[10] The input to a model is a sentence, and 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), and 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 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) and 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 Evaluation Framework

4.1 Pre-trained 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.

[9] https://www.sec.gov/edgar/

[10] Art. 3 of the Directive 93/13 on Unfair Terms in Consumer Contracts (http://data.europa.eu/eli/dir/1993/13/oj).

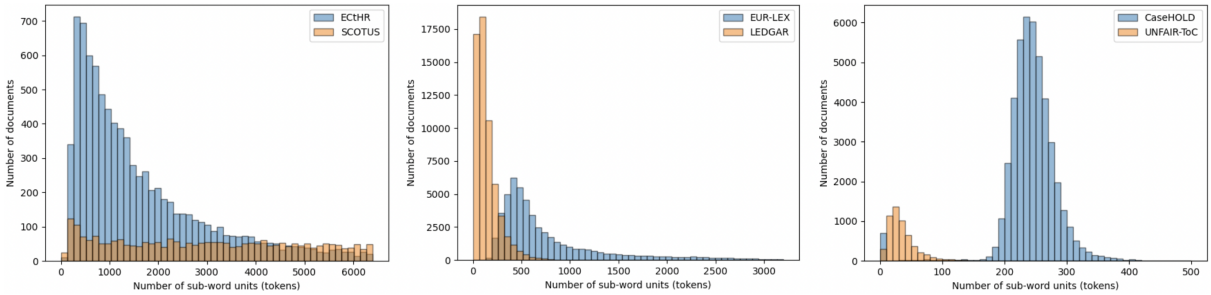

Figure 2: Distribution of text input length measured in BERT sub-word units across LexGLUE datasets.

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 much longer sequences using sparse-attention. The sparse-attention is a combination of a 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). 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). 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,[11] 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 highlight that it is trained exclusively on case law (court opinions).

4.2 Task-Specific Fine-Tuning

Hierarchical Model Legal documents are usually much longer (i.e. consisting of thousands of words) than typical text (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, ECtHR (A and B) and SCOTUS, we employ a hierarchical variant of all standard pre-trained Transformer-based models (BERT, RoBERTa, DeBERTa, Legal-BERT, CaseLaw-BERT) during fine-tuning and inference. The hierarchical variants are similar to that 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 $h_{\texttt{[cls]}}$ of each paragraph. A second-level shallow (2-layered) Transformer encoder with the exact same specifications (e.g., hidden units, number of

[11] https://case.law/

Method	ECtHR (A)*		ECtHR (B)*		SCOTUS*		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1/ m-F_1
BERT	**71.4**	64.0	79.6	**78.3**	70.5	60.9	71.6	55.6	87.7	82.2	97.3	80.4	70.7
RoBERTa	69.8	61.1	78.6	77.0	70.8	61.2	71.8	**57.5**	87.9	82.1	97.2	79.6	71.7
DeBERTa	69.3	63.0	79.9	**78.3**	70.0	60.0	**72.3**	57.2	87.9	82.0	97.2	80.2	72.1
Longformer	69.6	62.7	78.8	75.8	72.2	62.5	71.9	56.7	87.7	82.3	**97.5**	81.0	72.0
BigBird	69.7	62.2	79.9	76.9	71.7	61.4	71.8	56.6	87.7	82.1	97.4	81.1	70.4
Legal-BERT	71.2	**64.9**	**80.6**	77.2	76.2	65.8	72.2	56.2	**88.1**	**82.7**	97.4	**83.4**	75.1
CaseLaw-BERT	71.2	64.2	79.7	76.8	76.4	66.2	71.0	55.9	88.0	82.3	97.4	82.4	**75.6**

Table 3: Test results for all examined baselines across LexGLUE tasks. In datasets marked with an asterisk (*), we use the hierarchical variant of each model, as described in Section 4.1.

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, as in the rest of the datasets.[12]

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 $h_{\texttt{[cls]}}$ 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 model is employed, we feed the max-pooled 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 $h_{\texttt{[cls]}}$ of each pair is fed to a linear layer to obtain a logit (score), and the five logits are then passed through a softmax activation to obtain a probability distribution over the five candidate answers.

4.3 Experimental Set Up

For all the pre-trained models, we use publicly available Hugging Face checkpoints.[13] We use the *-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.[14] 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: 3x block, random: 3x block, global: 2x 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 the average scores across runs on the test set. We evaluate classification performance using *micro-F1* (μ-F_1) and *macro-F1* (m-F_1) across all datasets to take into account class imbalance.

4.4 Experimental Results

Table 3 presents the test results for all models across all LexGLUE tasks. We observe that the two legal-oriented pre-trained models (Legal-BERT, CaseLaw-BERT) perform overall better across all tasks. Their in-domain (legal) knowledge seems to be more critical in the two datasets relying on US case law data (SCOTUS, CaseHOLD) with an improvement of approximately +5% over the rest of the models, which are pre-trained on generic corpora. While these two models perform mostly on par, CaseLaw-BERT performs marginally better than Legal-BERT on SCOTUS and CaseHOLD, which is easily explained by the fact that it 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 legislation, ECtHR cases,

[12] In Appendix C, 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%).

[13] http://huggingface.co/models

[14] We omit results for *-large models, as we found them to be very unstable in preliminary experiments using fp16. See details in Appendix D.

US contracts; thus it performs slightly better in EUR-LEX, LEDGAR, UNFAIR-ToS, without performing better than CaseLaw-BERT, though, in the ECtHR tasks. No single model performs best in all tasks, and the results of Table 3 show that there is still large scope for improvements.

We note that we experimented exclusively with pre-trained Transformer models, because they currently achieve state of the art results in most NLP tasks. However, it would be interesting to experiment with simpler (and computationally much cheaper) models too (e.g., linear classifiers with TF-IDF features) to see how worse (if at all) they actually perform, and newcomers (e.g., students) could easily contribute results of this kind.

4.5 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).[15] Furthermore, we release the code of our experiments, which relies on the Hugging Face transformers (Wolf et al., 2020) library[16] on Github.[17] Details on how to load the datasets and how to run experiments with our code are available in Appendix A. We also provide a leaderboard where new results can be added.

5 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; Kitaev et al., 2020; Choromanski et al., 2021) have been proposed to handle long documents by exploring sparse attention. The publicly available models (Longformer, BigBird) can handle sequences up to 4096 sub-words, which is largely exceeded in three out of seven LexGLUE tasks (Figure 2).

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 hierarchical Transformer model that hierarchically encodes increasingly larger blocks (e.g., words, sentences, documents). SMITH is very similar to the hierarchical model of Section 4.1, 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 multi-jurisdictional 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). The effect of largely increasing model capacity has not been studied, however, in LexGLUE (the models we considered have up to 100M parameters). Hence, it is currently unclear what computing resources (and funding) one needs to excel in LexGLUE or, more generally, in domain-specific benchmarks.

6 Limitations and Future Work

In its current version, LexGLUE can only be used to evaluate English language 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 lack 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 schol-

[15] https://huggingface.co/datasets/lex_glue
[16] https://huggingface.co/transformers
[17] https://github.com/coastalcph/lex-glue

arly 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 in future work.

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.

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 (Angwin et al., 2016; Dressel and Farid, 2018), and robustness (Wang et al., 2021).

Acknowledgments

We would like to thank Desmond Elliott, Xiang Dai, and Joel Niklaus for providing valuable feedback. We are grateful to Cognitiv+ Ltd. for providing the compute infrastructure (8× Quadro RTX 6000 24GB GPU cards) for running the overwhelming number of experiments for all baselines.

References

Nikolaos Aletras, Ion Androutsopoulos, Leslie Barrett, Adam Meyers, and Daniel Preotiuc-Pietro, editors. 2020. *Proceedings of the 2nd Natural Legal Language Processing Workshop at KDD 2020*. Online.

Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, editors. 2019. *Proceedings of the 1st Natural Legal Language Processing Workshop at NAACL 2019*. Minneapolis, Minnesota.

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.

I. Angelidis, Ilias Chalkidis, and M. Koubarakis. 2018. Named entity recognition, linking and generation for greek legislation. In *JURIX*, Groningen, The Netherlands.

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*.

Pedro Henrique Luz de Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. 2020. VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

Purbid Bambroo and Aditi Awasthi. 2021. Legaldb: Long distilbert for legal document classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval*, pages 413–428, Cham. Springer International Publishing.

Reuben Binns. 2020. Analogies and disanalogies between machine-driven and human-driven legal judgement. *Journal of Cross-disciplinary Research in Computational Law*, 1(1).

Michael J. Bommarito, Daniel Martin Katz, and Eric M. Detterman. 2021. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *Research Handbook on Big Data Law*, pages 216–227.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik,

Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online.

L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29(2):213–238.

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.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Legal NERC with ontologies, Wikipedia and curriculum learning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 254–259, Valencia, Spain. Association for Computational Linguistics.

Ilias Chalkidis and Ion Androutsopoulos. 2017. A deep learning approach to contract element extraction. In *Proceedings of the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017)*, Luxembourg City, Luxembourg.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. 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, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 254–259, Melbourne, Australia. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019b. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy.

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 (EMNLP)*, Online.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020a. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online.

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

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019c. Neural contract element extraction revisited. In *Proceedings of the Document Intelligence Workshop at NeurIPS 2019*, Vancouver, Canada.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021b. Regulatory compliance through Doc2Doc information retrieval: A case study in EU/UK legislation where text similarity has limitations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3498–3511, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021c. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the Annual Conference of the North*

American Chapter of the Association for Computational Linguistics, online.

Ilias Chalkidis and Dimitrios Kampas. 2018. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571, online.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *International Conference on Learning Representations*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Corinna Coupette, Janis Beckedorf, Dirk Hartung, Michael Bommarito, and Daniel Martin Katz. 2021. Measuring law over time: A network analytical framework with an application to statutes and regulations in the United States and Germany. *Frontiers in Physics*, 9:269.

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.

Giuseppe Di Fatta, Victor Sheng, and Alfredo Cuzzocrea. 2020. The IEEE ICDM 2020 workshops. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 26–29. IEEE.

John S. Downie. 2004. IMIRSEL: a secure music retrieval testing environment. In *Internet Multimedia Management Systems V*, volume 5601, pages 91 – 99. International Society for Optics and Photonics, SPIE.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(10).

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186.

Rupert Haigh. 2018. *Legal English*. Routledge.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated NLP dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Abhik Jana and Chris Biemann. 2021. An investigation towards differentially private sequence tagging in a federated framework. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 30–35.

Yoshinobu Kano, Mi-Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of coliee 2017. In *COLIEE@ ICAIL*, pages 1–8.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2018. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 177–192. Springer.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Daniel Martin Katz, Corinna Coupette, Janis Beckedorf, and Dirk Hartung. 2020. Complex societies and the growth of the law. *Scientific Reports*, 10:18737.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998, online.

Mi-Young Kim, Randy Goebel, Yoshinobu Kano, and Ken Satoh. 2016. Coliee-2016: evaluation of the competition on legal information extraction and entailment. In *International Workshop on Juris-informatics (JURISIN 2016)*.

Mi-young Kim, Ying Xu, and Randy Goebel. 2015. A Convolutional Neural Network in Legal Question Answering. *Ninth International Workshop on Juris-informatics (JURISIN)*.

D. P. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Kankawin Kowsrihawat, Peerapon Vateekul, and Prachya Boonkwan. 2018. Predicting judicial decisions of criminal cases from thai supreme court using bi-directional gru with attention mechanism. In *2018 5th Asian Conference on Defense Technology (ACDT)*, pages 50–55. IEEE.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham. Springer International Publishing.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing.

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*, pages 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 preprint arXiv:1907.11692*.

Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, page 1261–1264, New York, NY, USA. Association for Computing Machinery.

Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 430–439, Vancouver, Canada. Association for Computational Linguistics.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: indian legal documents corpus for court judgmentprediction and explanation. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, online.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

Masha Medvedeva, Ahmet Üstun, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the European Court of Human Rights. In *Proceedings of the Fifth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021)*.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, 28(2):237–266.

Eneldo Loza Mencia and Johannes Fürnkranzand. 2007. An Evaluation of Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain. In *Proceedings of the 1st Linguistic Annotation Workshop*, pages 126–132, Halle, Germany.

Emre Mumcuoğlu, Ceyhun E Öztürk, Haldun M Ozaktas, and Aykut Koç. 2021. Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing & Management*, 58(5):102684.

Ramesh Nallapati and Christopher D. Manning. 2008. Legal docket classification: Where machine learning stumbles. In *Proceedings of the 2008 Conference on*

Empirical Methods in Natural Language Processing, pages 438–446, Honolulu, Hawaii. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Court-Predict: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the 3rd Natural Legal Language Processing Workshop Workshop*, Online.

Adam R Pah, David L Schwartz, Sarath Sanga, Zachary D Clopton, Peter DiCola, Rachel Davis Mersey, Charlotte S Alexander, Kristian J Hammond, and Luís A Nunes Amaral. 2020. How to build a more open justice system. *Science*, 369(6500):134–136.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China.

Theodore W Ruger, Pauline T Kim, Andrew D Martin, and Kevin M Quinn. 2004. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. *Columbia Law Review*, pages 1150–1210.

Yanchuan Sim, Bryan Routledge, and Noah A. Smith. 2016. Friends with motives: Using text to infer influence on SCOTUS. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1733, Austin, Texas. Association for Computational Linguistics.

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

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, pages 204–209.

Peter M Tiersma. 1999. *Legal language*. University of Chicago Press.

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 12th Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

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. SCITEPRESS - Science and Technology Publications.

Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, and Simone Teufel. 2021. What about the precedent: An information-theoretic analysis of common law. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2275–2288, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, California, USA.

Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Avinante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 130–135. IEEE.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Net-*

works for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, New Orleans, Louisiana, USA.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *Science China - Information Sciences*.

Christopher Williams. 2007. *Tradition and change in legal English: Verbal constructions in prescriptive texts*, volume 20. Peter Lang.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *CoRR*, abs/2105.03887.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. *Beyond 512 Tokens: Siamese Multi-Depth Transformer-Based Hierarchical Encoder for Long-Form Document Matching*, page 1725–1734. Association for Computing Machinery, New York, NY, USA.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4085–4091. International Joint Conferences on Artificial Intelligence Organization.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, pages 17283–17297, online.

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. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1250–1257.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020c. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 9701–9708, New York, NY, USA.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

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:

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

How to run experiments? To make reproducing the results of the already examined models or future models even easier, we release our

code on GitHub (https://github.com/coastalcph/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 randoms seeds, as we did for the reported results for the original leaderboard.

B No labeling as an additional class

In ECtHR Task A and B and UNFAIR-ToS, there are unlabeled samples. In ECtHR Task A, *no violation*, i.e., the court ruled that the defendant did not violate any ECHR article, is a possible event. Contrary, no violation is not a possible event in ECtHR Task B, i.e., at least a single ECHR article is allegedly violated (considered by the court), although there is such a rare scenario after simplifications, i.e., some cases were originally labeled only with rare labels, excluded in our benchmark (Section 3.2). Similarly in UNFAIR-ToS, the vast majority of sentences are not labeled with any type of *unfairness* (unfair term against users), i.e., the sentence does not raise any question for the violation of the European consumer law.

In multi-label (multi-class multi-output) classification, the set of labels per example is represented as a one-hot vector ($Y = [y_1, y_2, \ldots, y_L]$), where y_i represents the ith class and whose value is zero (0) if the example is not labeled with the specific class and one (1), otherwise. In case the example is unlabeled (the example is not labeled with any class), the one-hot vector includes only zeros. During training, cross-entropy loss is effectively affected (informed) by such examples, as the predictions ($\hat{Y} = [\hat{y}_1, \hat{y}_2 \ldots, \hat{y}_L]$) diverge from zeros. Although during evaluation, F1-score does not consider such examples when both $Y = \hat{Y} = \{i : y_i = 0\}$. Hence, in order to consider the correct labeling of such examples in F1-score, during evaluation, we include an additional label (y_0) in both targets (Y), and predictions ($\hat{Y}$), whose value is 1 (positive), if the value is 0 (negative) across all other labels, 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).

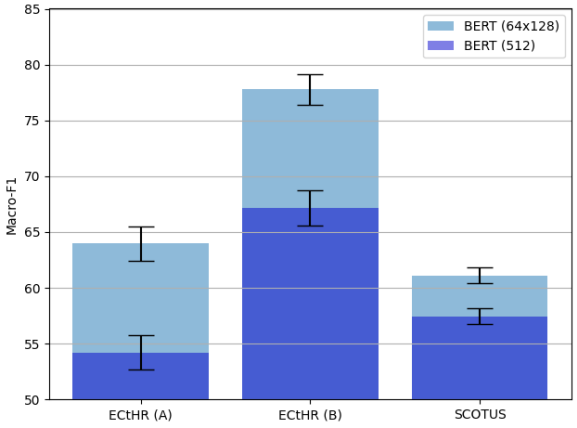

Figure 3: Results of BERT models in ECtHR (Task A and B) and SCOTUS. Light blue denotes the average score across 5 runs for the hierarchical variant of BERT, presented in Table 3, while dark blue corresponds to the standard BERT model (up to 512 tokens). The black error bars show the standard error.

C Use of standard BERT models

In Table 3, we present the results for the standard BERT model of Devlin et al. (2019) using up to 512 tokens, compared to its hierarchical variant. We observe that across all datasets, where documents are long (Figure 2(a)), the hierarchical variant clearly outperforms the standard model fed with truncated documents (ECtHR A: +12.2%, ECtHR B: 10.6%, SCOTUS: 3.5%). Compared to the ECtHR tasks, the gains are lower in SCOTUS, a topic classification task where long-range reasoning is not needed, i.e., in contrast for ECtHR, multiple distant facts need to be combined.

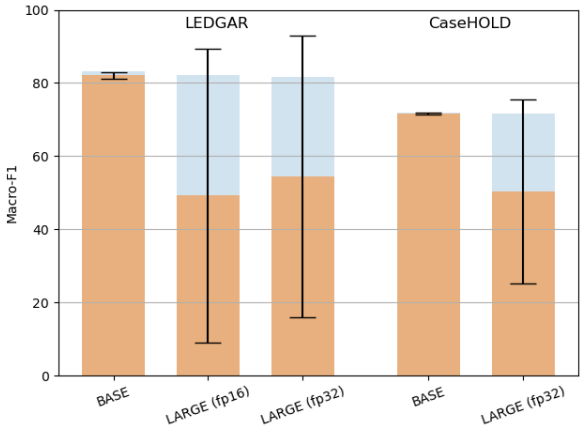

Figure 4: Results of RoBERTa models in LEDGAR and CaseHOLD. Blue and orange denote the maximum and average scores, respectively, across 5 runs. The black error bars show the standard error.

Method	ECtHR (A)*		ECtHR (B)*		SCOTUS*		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1/ m-F_1
BERT	70.6	64.1	79.8	**80.3**	76.6	68.8	77.2	61.1	87.9	81.8	**97.3**	76.2	72.6
RoBERTa	69.8	64.3	79.5	79.9	76.2	68.4	77.5	**63.1**	87.9	81.4	96.8	74.4	73.9
DeBERTa	69.2	62.9	79.9	**80.3**	77.7	63.1	**77.7**	**63.1**	88.1	81.5	97.2	76.0	73.8
Longformer	70.8	64.4	**80.8**	76.5	76.9	70.1	77.4	62.5	88.1	81.7	97.2	74.3	73.7
BigBird	70.9	65.0	79.5	79.3	75.9	68.7	77.2	62.1	87.9	81.5	97.2	**78.0**	73.5
Legal-BERT	71.8	**67.8**	80.7	79.7	80.1	72.3	77.5	61.5	**88.3**	**81.9**	97.2	**78.0**	76.1
CaseLaw-BERT	**71.9**	66.3	80.2	78.2	**81.1**	**73.4**	77.2	62.3	**88.3**	**81.9**	**97.3**	76.6	**77.2**

Table 4: Validation results for all examined baselines across LexGLUE tasks. In datasets marked with an asterisk (*), we use the hierarchical variant of each model, as described in Section 4.1.

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

Table 5: Training time in total (T) and per epoch (T/e) across LexGLUE tasks. In datasets marked with asterisk (*), we use the hierarchical variant of each model, as described in Section 4.1.

D Use of Larger Models

In preliminary experiments, we also considered pre-trained models using the *-large configuration, i.e., 24 Transformer blocks, 1024 units, and 18 attentions heads, namely RoBERTa-large Liu et al. (2019). Similarly to the rest of the experiments, we used mixed precision (fp16) to decrease the memory footprint in training. Using fp16 in attention operation resulted in floating point overflow and NaNs in later stages of training. Similar issues have been also reported by Beltagy et al. (2020). In our case, we face similar issues even training models with the standard fp32 setting. In Table 4, we present the results across runs for RoBERTa-large.

E Additional Results

For completeness, Table 4 shows the validation results for all examined models across all datasets. Results are improved and also vary comparing to those reported for the test set, which further highlights the importance of temporal splits. In Table 5, we present information on the training time per dataset and model.

F Other Tasks/Datasets Considered

We primarily 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, etc.). 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.[18] The original dataset follows the SQUAD v2.0 setting, including unanswerable questions. Following the SQUAD v1.1 (Rajpurkar et al., 2016) setting, we simplified the task by removing all unanswerable pairs (question, paragraph). The information in the original dataset is very sparse leading to a vast majority of unanswerable pairs. We also excluded all annotations that 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) 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 probabil-

[18] The question mostly resembles a *prompt*, rather than a natural question, as there is a closed set of 41 alternatives.

ity 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, similar 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).[19] 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 these 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). In this case (Parties), those mentions are actually party roles in the context of a contract.

- Annotations may include semi-redacted text (e.g., "_____, 1996" for *Agreement Date*), or even fully redacted text (e.g., "_____" for *Parties*). This practice secures () information in public filings, but the scope of the examined task is to process everyday business contracts, hence 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, underestimating 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 do plan, however, to include more demanding tasks in future LexGLUE versions, as the legal NLP community will be growing.

[19] F1 is one of the two official SQUAD measures. In the second one, Exact Answer Accuracy, all models scored 0%.