LEGAL-BERT: The Muppets straight out of Law School

Ilias Chalkidis [†‡] **Manos Fergadiotis** [†‡]
Prodromos Malakasiotis [†‡] **Nikolaos Aletras** [*] **Ion Androutsopoulos** [†‡]
[†] Department of Informatics, Athens University of Economics and Business
[‡] Institute of Informatics & Telecommunications, NCSR "Demokritos"
[*] Computer Science Department, University of Sheffield, UK
[ihalk,fergadiotis,rulller,ion]@aueb.gr
n.aletras@sheffield.ac.uk

Abstract

BERT has achieved impressive performance in several NLP tasks. However, there has been limited investigation on its adaptation guidelines in specialised domains. Here we focus on the legal domain, where we explore several approaches for applying BERT models to downstream legal tasks, evaluating on multiple datasets. Our findings indicate that the previous guidelines for pre-training and fine-tuning, often blindly followed, do not always generalize well in the legal domain. Thus we propose a systematic investigation of the available strategies when applying BERT in specialised domains. These are: (a) use the original BERT out of the box, (b) adapt BERT by additional pre-training on domain-specific corpora, and (c) pre-train BERT from scratch on domain-specific corpora. We also propose a broader hyper-parameter search space when fine-tuning for downstream tasks and we release LEGAL-BERT, a family of BERT models intended to assist legal NLP research, computational law, and legal technology applications.

1 Introduction

Pre-trained language models based on Transformers (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and its variants (Liu et al., 2019; Yang et al., 2019; Lan et al., 2019), have achieved state-of-the-art results in several downstream NLP tasks on generic benchmark datasets, such as GLUE (Wang et al., 2018), SQUAD (Rajpurkar et al., 2016), and RACE (Lai et al., 2017).

Typically, transfer learning with language models requires a computationally heavy step where the language model is pre-trained on a large corpus and a less expensive step where the model is fine-tuned for downstream tasks. When using BERT, the first step can be omitted as the pre-trained models are publicly available. Being pre-trained on generic corpora (e.g., Wikipedia, Children's Books, etc.) BERT has been reported to under-perform in specialised domains, such as biomedical or scientific text (Lee et al., 2019; Beltagy et al., 2019). To overcome this limitation there are two possible strategies; either further pre-train (FP) BERT on domain specific corpora, or pre-train BERT from scratch (SC) on domain specific corpora. Consequently, to employ BERT in specialised domains one may consider three alternative strategies before fine-tuning for the downstream task (Figure 1): (a) use BERT out of the box, (b) further pre-train (FP) BERT on domain-specific corpora, and (c) pre-train BERT from scratch (SC) on domain specific corpora with a new vocabulary of sub-word units.

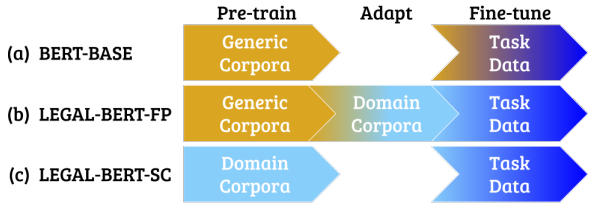

Figure 1: The three alternatives when employing BERT for NLP tasks in specialised domains: (a) use BERT out of the box, (b) further pre-train BERT (FP), and (c) pre-train BERT from scratch (SC). All strategies have a final fine-tuning step.

In this paper, we systematically explore strategies (a)–(c) in the legal domain, where BERT adaptation has yet to be explored. As with other specialised domains, legal text (e.g., laws, court pleadings, contracts) has distinct characteristics compared to generic corpora, such as specialised vocabulary, particularly formal syntax, semantics based on extensive domain-specific knowledge etc., to the extent that legal language is often classified as a 'sublanguage' (Tiersma, 1999; Williams, 2007; Haigh, 2018). Note, however, that our work contributes more broadly towards a better understanding of domain adaptation for specialised domains. Our key findings are: (i) Further pre-training (FP) or pre-training BERT from scratch (SC) on domain-

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2898–2904

Corpus	No. documents	Total Size in GB	Repository
EU legislation	61,826	1.9 (16.5%)	EURLEX (eur-lex.europa.eu)
UK legislation	19,867	1.4 (12.2%)	LEGISLATION.GOV.UK (http://www.legislation.gov.uk)
European Court of Justice (ECJ) cases	19,867	0.6 (5.2%)	EURLEX (eur-lex.europa.eu)
European Court of Human Rights (ECHR) cases	12,554	0.5 (4.3%)	HUDOC (http://hudoc.echr.coe.int)
US court cases	164,141	3.2 (27.8%)	CASE LAW ACCESS PROJECT (https://case.law)
US contracts	76,366	3.9 (34.0%)	SEC-EDGAR (https://www.sec.gov/edgar.shtml)

Table 1: Details on the training corpora used to pre-train the different variations of LEGAL-BERT. All repositories have open access, except from the Case Law Access Project, where access is granted to researchers upon request.

specific corpora, performs better than using BERT out of the box for domain-specific tasks; both strategies are mostly comparable in three legal datasets. (ii) Exploring a broader hyper-parameter range, compared to the guidelines of Devlin et al. (2019), can lead to substantially better performance. (iii) Smaller BERT-based models can be competitive to larger, computationally heavier ones in specialised domains. Most importantly, (iv) we release LEGAL-BERT, a family of BERT models for the legal domain, intended to assist legal NLP research, computational law, and legal technology applications.[1] This family includes LEGAL-BERT-SMALL, a light-weight model pre-trained from scratch on legal data, which achieves comparable performance to larger models, while being much more efficient (approximately 4 times faster) with a smaller environmental footprint (Strubell et al., 2019).

2 Related Work

Most previous work on the domain-adaptation of BERT and variants does not systematically explore the full range of the above strategies and mainly targets the biomedical or broader scientific domains. Lee et al. (2019) studied the effect of further pre-training BERT-BASE on biomedical articles for 470k steps. The resulting model (BIOBERT) was evaluated on biomedical datasets, reporting performance improvements compared to BERT-BASE. Increasing the additional domain-specific pre-training to 1M steps, however, did not lead to any clear further improvements. Alsentzer et al. (2019) released Clinical BERT and Clinical BIOBERT by further pre-training BERT-BASE and BIOBERT, respectively, on clinical notes for 150k steps. Both models were reported to outperform BERT-BASE. In other related work, Beltagy et al. (2019) released SCIBERT, a family of BERT-based models for scientific text, with emphasis on the biomedical domain. Their models were obtained either by further pre-training (FP) BERT-BASE, or by pre-training BERT-BASE from scratch (SC) on a domain-specific corpus, i.e., the model is randomly initialized and the vocabulary was created from scratch. Improvements were reported in downstream tasks in both cases. Sung et al. (2019) further pre-trained BERT-BASE on textbooks and question-answer pairs to improve short answer grading for intelligent tutoring systems.

One shortcoming is that all previous work does not investigate the effect of varying the number of pre-training steps, with the exception of Lee et al. (2019). More importantly, when fine-tuning for the downstream task, all previous work blindly adopts the hyper-parameter selection guidelines of Devlin et al. (2019) without further investigation. Finally, no previous work considers the effectiveness and efficiency of smaller models (e.g., fewer layers) in specialised domains. The full capacity of larger and computationally more expensive models may be unnecessary in specialised domains, where syntax may be more standardized, the range of topics discussed may be narrower, terms may have fewer senses etc. We also note that although BERT is the current state-of-the-art in many legal NLP tasks (Chalkidis et al., 2019c,a,d), no previous work has considered its adaptation for the legal domain.

3 LEGAL-BERT: A new family of BERT models for the legal domain

Training corpora: To pre-train the different variations of LEGAL-BERT, we collected 12 GB of diverse English legal text from several fields (e.g., legislation, court cases, contracts) scraped from publicly available resources (see Table 1).

LEGAL-BERT-FP: Following Devlin et al. (2019), we run additional pre-training steps of BERT-BASE on domain-specific corpora. While Devlin et al. (2019) suggested additional steps up to 100k, we also pre-train models up to 500k to examine the effect of prolonged in-domain pre-training when fine-tuning on downstream tasks. BERT-BASE has been pre-trained for significantly more steps in generic corpora (e.g., Wikipedia, Children's Books), thus it

[1] All models and code examples are available at: https://huggingface.co/nlpaueb.

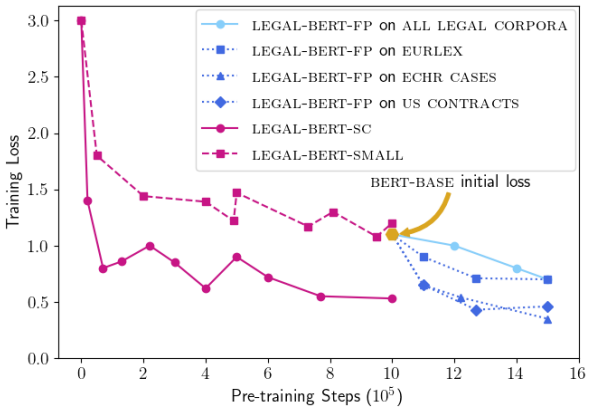

Figure 2: Train losses for all LEGAL-BERT versions.

is highly skewed towards generic language, using a vocabulary of 30k sub-words that better suits these generic corpora. Nonetheless we expect that prolonged in-domain pre-training will be beneficial.

LEGAL-BERT-SC has the same architecture as BERT-BASE with 12 layers, 768 hidden units and 12 attention heads (110M parameters). We use this architecture in all our experiments unless otherwise stated. We use a newly created vocabulary of equal size to BERT's vocabulary.[2] We also experiment with LEGAL-BERT-SMALL, a substantially smaller model, with 6 layers, 512 hidden units, and 8 attention heads (35M parameters, 32% the size of BERT-BASE). This light-weight model, trains approx. 4 times faster, while also requiring fewer hardware resources.[3] Our hypothesis is that such a specialised BERT model can perform well against generic BERT models, despite its fewer parameters.

4 Experimental Setup

Pre-training Details: To be comparable with BERT, we train LEGAL-BERT for 1M steps (approx. 40 epochs) over all corpora (Section 3), in batches of 256 samples, including up to 512 sentencepiece tokens. We used Adam with learning rate of $1e{-}4$, as in the original implementation. We trained all models with the official BERT code[4] using v3 TPUs with 8 cores from Google Cloud Compute Services.

Legal NLP Tasks: We evaluate our models on text classification and sequence tagging using three datasets. EURLEX57K (Chalkidis et al., 2019b) is a large-scale multi-label text classification dataset of EU laws, also suitable for few and zero-shot learning. ECHR-CASES (Chalkidis et al., 2019a) contains cases from the European Court of Human Rights (Aletras et al., 2016) and can be used for binary and multi-label text classification. Finally, CONTRACTS-NER (Chalkidis et al., 2017, 2019d) is a dataset for named entity recognition on US contracts consisting of three subsets, *contract header*, *dispute resolution*, and *lease details*. We replicate the experiments of Chalkidis et al. (2019c,a,d) when fine-tuning BERT for all datasets.[5]

Tune your Muppets! As a rule of thumb to fine-tune BERT for downstream tasks, Devlin et al. (2019) suggested a minimal hyper-parameter tuning strategy relying on a grid-search on the following ranges: learning rate $\in \{2e{-}5, 3e{-}5, 4e{-}5, 5e{-}5\}$, number of training epochs $\in \{3, 4\}$, batch size $\in \{16, 32\}$ and fixed dropout rate of 0.1. These not well justified suggestions are blindly followed in the literature (Lee et al., 2019; Alsentzer et al., 2019; Beltagy et al., 2019; Sung et al., 2019). Given the relatively small size of the datasets, we use batch sizes $\in \{4, 8, 16, 32\}$. Interestingly, in preliminary experiments, we found that some models still underfit after 4 epochs, the maximum suggested, thus we use early stopping based on validation loss, without a fixed maximum number of training epochs. We also consider an additional lower learning rate ($1e{-}5$) to avoid overshooting local minima, and an additional higher drop-out rate (0.2) to improve regularization. Figure 4 (top two bars) shows that our enriched grid-search (*tuned*) has a substantial impact in most of the end-tasks compared to the default hyper-parameter strategy of Devlin et al. (2019).[6] We adopt this strategy for LEGAL-BERT.

5 Experimental Results

Pre-training Results: Figure 2 presents the training loss across pre-training steps for all versions of LEGAL-BERT. LEGAL-BERT-SC performs much better on the pre-training objectives than LEGAL-BERT-SMALL, which was highly expected, given the different sizes of the two models. At the end of its pre-training, LEGAL-BERT-SMALL has similar loss to that of BERT-BASE pre-trained on generic corpora (arrow in Figure 2). When we consider the additional pre-training of BERT on legal corpora

[2] We use Google's sentencepiece library (https://github.com/google/sentencepiece.)

[3] Consult Appendix C for a comparison on hardware resources as well as training and inference times.

[4] github.com/google-research/bert

[5] For implementation details, see Appendices A and B.

[6] In the *lease details* subset of CONTRACTS-NER, the optimal hyper-parameters fall in the ranges of Devlin et al. (2019).

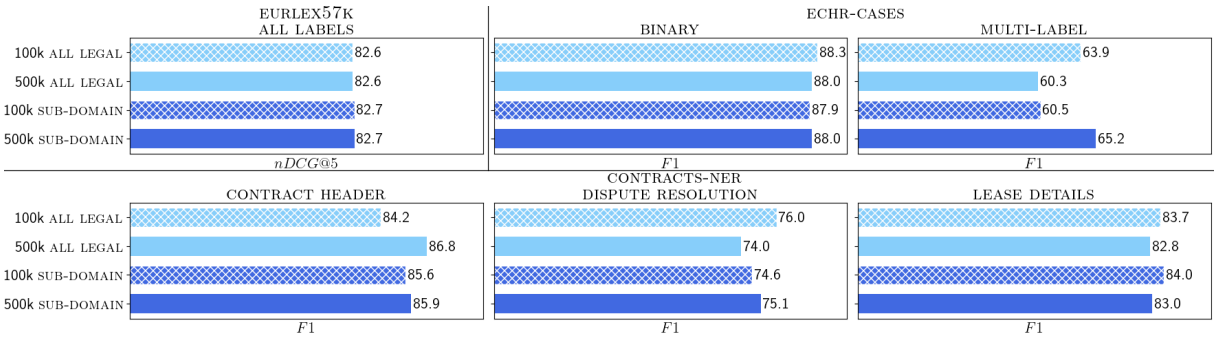

Figure 3: End-task results on development data across all datasets for LEGAL-BERT-FP variants.

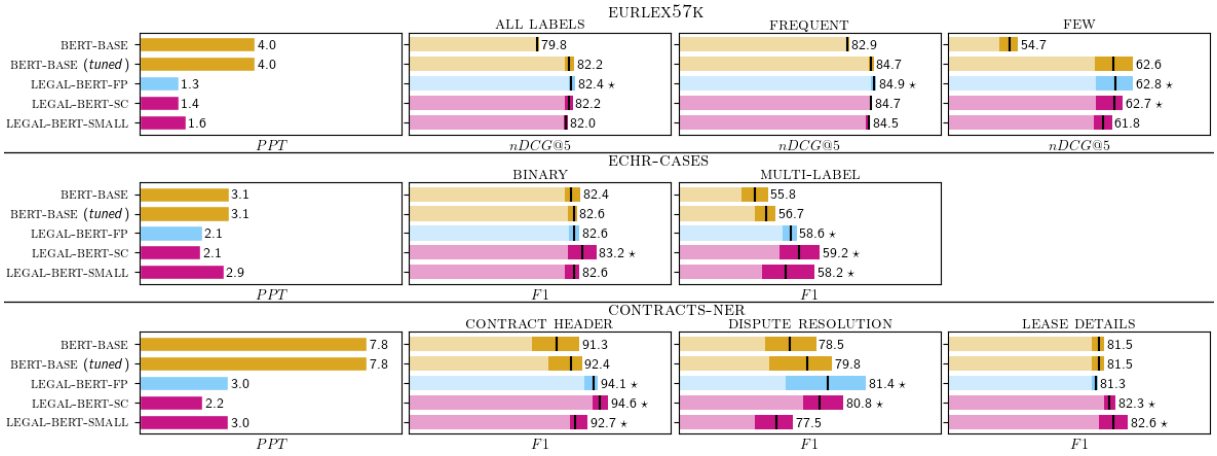

Figure 4: Perplexities (*PPT*) and end-task results on test data across all datasets and all models considered. The reported results are averages over multiple runs also indicated by a vertical black line in each bar. The transparent and opaque parts of each bar show the minimum and maximum scores of the runs, respectively. A star indicates versions of LEGAL-BERT that perform better on average than the tuned BERT-BASE.

(LEGAL-BERT-FP), we observe that it adapts faster and better in specific sub-domains (esp. ECHR cases, US contracts), comparing to using the full collection of legal corpora, where the training loss does not reach that of LEGAL-BERT-SC.

End-task Results: Figure 3 presents the results of all LEGAL-BERT-FP variants on development data. The optimal strategy for further pre-training varies across datasets. Thus in subsequent experiments on test data, we keep for each end-task the variant of LEGAL-BERT-FP with the best development results.

Figure 4 shows the perplexities and end-task results (minimum, maximum, and averages over multiple runs) of all BERT variants considered, now on test data. Perplexity indicates to what extent a BERT variant predicts the language of an end-task. We expect models with similar perplexities to also have similar performance. In all three datasets, a LEGAL-BERT variant almost always leads to better results than the tuned BERT-BASE. In EURLEX57K, the improvements are less substantial for *all*, *frequent*, and *few* labels (0.2%), also in agreement with the small drop in perplexity (2.7). In ECHR-CASES, we again observe small differences in perplexities (1.1 drop) and in the performance on the binary classification task (0.8% improvement). On the contrary, we observe a more substantial improvement in the more difficult multi-label task (2.5%) indicating that the LEGAL-BERT variations benefit from in-domain knowledge. On CONTRACTS-NER, the drop in perplexity is larger (5.6), which is reflected in the increase in $F1$ on the *contract header* (1.8%) and *dispute resolution* (1.6%) subsets. In the *lease details* subset, we also observe an improvement (1.1%). Impressively, LEGAL-BERT-SMALL is comparable to LEGAL-BERT across most datasets, while it can fit in most modern GPU cards. This is important for researchers and practitioners with limited access to large computational resources. It also provides a more memory-friendly basis for more complex BERT-based architectures. For example, deploying a hierarchical version of BERT for ECHR-CASES (Chalkidis et al., 2019a) leads to a $4\times$ memory increase.

6 Conclusions and Future Work

We showed that the best strategy to port BERT to a new domain may vary, and one may consider either further pre-training or pre-training from scratch. Thus, we release LEGAL-BERT, a family of BERT models for the legal domain achieving state-of-art results in three end-tasks. Notably, the performance gains are stronger in the most challenging end-tasks (i.e., *multi-label* classification in ECHR-CASES and *contract header*, *lease details* in CONTRACTS-NER) where in-domain knowledge is more important. We also release LEGAL-BERT-SMALL, which is 3 times smaller but highly competitive to the other versions of LEGAL-BERT. Thus, it can be adopted more easily in low-resource test-beds. Finally, we show that an expanded grid search when fine-tuning BERT for end-tasks has a drastic impact on performance and thus should always be adopted. In future work, we plan to explore the performance of LEGAL-BERT in more legal datasets and tasks. We also intend to explore the impact of further pre-training LEGAL-BERT-SC and LEGAL-BERT-SMALL on specific legal sub-domains (e.g., EU legislation).

Acknowledgments

This project was supported by the Google Cloud Compute (GCP) research program, while we also used a Google Cloud TPU v3-8 for free provided by the TensorFlow Research Cloud (TFRC) program[7]. We are grateful to both Google programs.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611, Hong Kong, China.

I. Chalkidis, I. Androutsopoulos, and A. Michos. 2017. Extracting Contract Elements. In *Proceedings of the International Conference of AI and Law*, London, UK.

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.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019b. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019c. 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, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019d. Neural Contract Element Extraction Revisited. In *Proceedings of the Document Intellgence Workshop collocated with NeurIPS 2019*, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, abs/1810.04805.

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

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *CoRR*, abs/1909.11942.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In *CoRR*.

[7] https://www.tensorflow.org/tfrc

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. *CoRR*, abs/1907.11692.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *ArXiv*, abs/2002.12327.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training BERT on domain resources for short answer grading. 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 6071–6075, Hong Kong, China. Association for Computational Linguistics.

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

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 *31th Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA.

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

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR*, abs/1906.08237.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

A Legal NLP datasets

Bellow are the details of the legal NLP datasets we used for the evaluation of our models:

- EURLEX57K (Chalkidis et al., 2019b) contains 57k legislative documents from EURLEX with an average length of 727 words. All documents have been annotated by the Publications Office of EU with concepts from EUROVOC.[8] The average number of labels per document is approx. 5, while many of them are rare. The dataset is split into *training* (45k), *development* (6k), and *test* (6k) documents.

- ECHR-CASES (Chalkidis et al., 2019a) contains approx. 11.5k cases from ECHR's public database. For each case, the dataset provides a list of *facts*. Each case is also mapped to *articles* of the Human Rights Convention that were violated (if any). The dataset can be used for binary classification, where the task is to identify if there was a violation or not, and for multi-label classification where the task is to identify the violated articles.

- CONTRACTS-NER (Chalkidis et al., 2017, 2019d) contains approx. 2k US contracts from EDGAR. Each contract has been annotated with multiple contract elements such as *title*, *parties*, *dates of interest*, *governing law*, *jurisdiction*, *amounts* and *locations*, which have been organized in three groups (*contract header*, *dispute resolution*, *lease details*) based on their position in contracts.

B Implementation details and results on downstream tasks

Below we describe the implementation details for fine-tuning BERT and LEGAL-BERT on the three downstream tasks:

EURLEX57K: We replicate the experiments of Chalkidis et al. (2019c), where a linear layer with L (number of labels) sigmoid activations was placed on top of BERT's `[CLS]` final representation. We follow the same configuration for all LEGAL-BERT variations.

[8] http://eurovoc.europa.eu/

ECHR-CASES: We replicate the best method of Chalkidis et al. (2019a), which is a hierarchical version of BERT, where initially a shared BERT encodes each case fact independently and produces N fact embeddings (`[CLS]` representations). A self-attention mechanism, similar to Yang et al. (2016), produces the final document representation. A linear layer with $\mathrm{softmax}$ activation gives the final scores.

CONTRACTS-NER We replicate the experiments of Chalkidis et al. (2019d) in all of their three parts (*contract header*, *dispute resolution*, *lease details*). In these experiments, the final representations of the original BERT for all (sentencepiece) tokens in the sequence are fed to a linear CRF layer.

We again follow Chalkidis et al. (2019c,a,d) in the reported evaluation measures.

C Efficiency comparison for various BERT-based models

Model.	Params	T	HU	AH	Max BS	Training Speed		Inference Speed
						$BS = 1$	$BS = \max$	$BS = 1$
BERT-BASE	110M	12	768	12	6	1.00×	1.00×	1.00×
ALBERT.	12M	12	768	12	12	1.26×	1.21×	1.00×
ALBERT-LARGE	18M	24	1024	12	4	0.49×	0.37×	0.36×
DISTIL-BERT	66M	6	768	12	16	1.66×	2.36×	1.70×
LEGAL-BERT	110M	12	768	12	6	1.00×	1.00×	1.00×
LEGAL-BERT-SMALL	35M	6	512	8	26	2.43×	4.00×	1.70×

Table 2: Comparison of BERT-based models for different batch sizes (BS) in a single 11GB NVIDIA-2080TI. Resource efficiency of the models mostly relies on the number of hidden units (HU), attentions heads (AH) and Transformer blocks T, rather than the number of parameters.

Recently there has been a debate on the over-parameterization of BERT (Kitaev et al., 2020; Rogers et al., 2020). Towards that directions most studies suggest a parameter sharing technique (Lan et al., 2019) or distillation of BERT by decreasing the number of layers (Sanh et al., 2019). However the main bottleneck of transformers in modern hardware is not primarily the total number of parameters, misinterpreted into the number of stacked layers. Instead Out Of Memory (OOM) issues mainly happen as a product of wider models in terms of hidden units' dimensionality and the number of attention heads, which affects gradient accumulation in feed-forward and multi-head attention layers (see Table 2). Table 2 shows that LEGAL-BERT-SMALL despite having $3\times$ and $2\times$ the parameters of ALBERT and ALBERT-LARGE has faster training and inference times. We expect models overcoming such limitations to be widely adopted by researchers and practitioners with limited resources. Towards the same direction Google released several lightweight versions of BERT.[9]

[9] https://github.com/google-research/bert