LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development

Ilias Chalkidis* **Nicolas Garneau*** **Anders Søgaard**
Department of Computer Science, University of Copenhagen, Denmark

Cătălina Goanță
Utrecht University School of Law, Netherlands

Daniel Martin Katz
Illinois Tech – Chicago Kent College of Law, IL, United States

Abstract

In this work, we conduct a detailed analysis on the performance of legal-oriented pre-trained language models (PLMs). We examine the interplay between their original objective, acquired knowledge, and legal language understanding capacities which we define as the upstream, probing, and downstream performance, respectively. We consider not only the models' size but also the pre-training corpora used as important dimensions in our study. To this end, we release a multinational English legal corpus (LeXFiles) and a legal knowledge probing benchmark (LegalLAMA) to facilitate training and detailed analysis of legal-oriented PLMs. We release two new legal PLMs trained on LeXFiles and evaluate them alongside others on LegalLAMA and LexGLUE. We find that probing performance strongly correlates with upstream performance in related legal topics. On the other hand, downstream performance is mainly driven by the model's size and prior legal knowledge which can be estimated by upstream and probing performance. Based on these findings, we can conclude that both dimensions are important for those seeking the development of domain-specific PLMs.

1 Introduction

Following closely the advances in the development of NLP technologies, the legal NLP literature is flourishing with the release of many new resources, including large legal corpora (Henderson* et al., 2022), datasets (Chalkidis et al., 2021a; Koreeda and Manning, 2021; Zheng et al., 2021; Chalkidis et al., 2022a; Habernal et al., 2022), and pre-trained legal-oriented language models (PLMs) (Chalkidis et al., 2020; Zheng et al., 2021; Xiao et al., 2021). Benchmark suites (Chalkidis et al., 2022a; Hwang et al., 2022; Niklaus et al., 2023) to evaluate the performance of PLMs in a more systematic way have been also developed, showcasing the superiority of legal-oriented PLMs over generic ones on downstream legal NLP tasks.

Despite this impressive progress, there is still not a thorough study on (a) how PLMs trained under different settings (pre-training corpora, size of the model) perform across different legal sub-corpora, and (b) what sort of knowledge such models have acquired from pre-training, and (c) how important is domain (legal) specificity vs general (cross-domain) legal knowledge. Furthermore, often times, legal NLP relies on datasets without drawing clear lines and comparisons between the various legal systems they may reflect. A legal system may be defined as a set of rules adopted and enforced at a given governance level, which may be national, regional or international (Friedman and Hayden, 2017), e.g., UK, EU, US, CoE, etc.

We define the upstream evaluation as the task PLMs are explicitly designed to do: Masked Language Modelling (MLM) (Devlin et al., 2019). We then probe for specific legal concepts that are legal-system specific, in a similar fashion as Petroni et al. (2019) did using the "LAnguage Models Analysis" (LAMA) framework. Finally, we assess the PLMs performance in LexGLUE (Chalkidis et al., 2022a) downstream tasks. More importantly, we explore how the aforementioned factors (upstream, and probing performance) interplay and relate to downstream performance. Our contributions are:

(a) We release LeXFiles, a new diverse English legal corpus including 11 sub-corpora that cover legislation and case law from 6 primarily English-speaking legal systems (EU, CoE, Canada, US, UK, India). The corpus comprises approx. 6 million documents which sum up to approx. 19 billion tokens.

(b) We release 2 new legal-oriented PLMs, dubbed LexLMs, warm-started from the RoBERTa (Liu et al., 2019) models, and further pre-trained on the LeXFiles for 1M additional steps.

* Equal contribution.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15513–15535
July 9-14, 2023

Sub-Corpus (Source)	# Documents	# Tokens / Percentage (%)	Sampling Smoothing (%)
EU Legislation	93.7K	233.7M (01.2%)	05.0%
EU Case Law	29.8K	178.5M (00.9%)	04.3%
UK Legislation	52.5K	143.6M (00.7%)	03.9%
UK Case Law	47K	368.4M (01.9%)	06.2%
Canadian Legislation	6K	33.5M (00.2%)	01.9%
Canadian Case Law	11.3K	33.1M (00.2%)	01.8%
U.S. Legislation	518	1.4B (07.4%)	12.3%
U.S. Case Law	4.6M	11.4B (59.2%)	34.7%
U.S. Contracts	622K	5.3B (27.3%)	23.6%
ECtHR Case Law	12.5K	78.5M (00.4%)	02.9%
Indian Case Law	34.8K	111.6M (00.6%)	03.4%
Total	5.8M	18.8B (100%)	100%

Table 1: Core statistics of the newly introduced LEXFILES corpus. In the last column, we present the sampling smoothing percentages used to train our LexLM models (Section 4.1).

(c) We release LEGALLAMA, a diverse probing benchmark suite comprising 8 sub-tasks that aims to assess the acquaintance of legal knowledge that PLMs acquired in pre-training.

(d) We evaluate 7 PLMs on both LEXFILES and LEGALLAMA, analyzing their performance out of the box per LEXFILES sub-corpus and LEGALLAMA tasks. We also fine-tune and evaluate these models in selected LEXGLUE tasks, and examine the interplay between MLM, probing, and downstream performance.

2 LeXFiles Corpus

The LEXFILES is a new diverse English multinational legal corpus that we created including 11 distinct sub-corpora (Table 1) that cover legislation and case law from 6 primarily English-speaking legal systems (EU, CoE, Canada, US, UK, India). The corpus contains approx. 19 billion tokens. In comparison, the PILE OF LAW corpus released by Henderson* et al. (2022) comprises 32 billion in total, where the majority (26/30) of sub-corpora come from the United States of America (USA), hence the corpus as a whole is biased towards the US legal system in general, and the federal or state jurisdiction in particular, to a significant extent. The LEXFILES's sub-corpora are:

(a) *EU Legislation.* We release 93.7K EU laws (regulations, decisions, directives) published in EUR-Lex, the website of the EU Publication Office.[1]

(b) *EU Case Law.* We release 29.8K EU court decisions, mainly issued from the Court of Justice (CJEU), published in EUR-Lex.[1]

(c) *UK Legislation.* We release 52.5 UK laws published in UK.LEGISLATION.GOV.UK, the official website of the UK National Archives.[2]

(d) *UK Case Law.* We release 47K UK court decisions published in the British and Irish Legal Information Institute (BAILII) database.[3]

(e) *US Legislation.* We re-distribute 518 US state statutes (legislation) originally published by Henderson* et al. (2022).

(f) *US Case Law.* We release 4.6M US decisions (opinions) published by Court Listener,[4] a web database hosted by the Free Law Project.[5]

(g) *US Contracts.* We release 622K US contracts (agreements) obtained from US Securities and Exchange Commission (SEC) filings, which are publicly available from the SEC-EDGAR[6] database.

(h) *Canadian Legislation.* We release 6K Canadian laws (acts, regulations) published in the official legislation portal of Canada.[7]

(i) *Canadian Case Law.* We re-distribute 13.5K Canadian decisions (opinions) originally published by Henderson* et al. (2022).

(j) *ECtHR Case Law.* We release 12.5K decisions ruled by the European Court of Human rights

[1] https://eur-lex.europa.eu/

[2] https://www.legislation.gov.uk/

[3] https://www.bailii.org/

[4] https://www.courtlistener.com/

[5] We release decisions published from 1965 on-wards (cf. post Civil Rights Act), as a hard threshold for cases that possibly rely on out-dated and discriminatory law standards. The rest of the sub-corpora include more recent documents.

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

[7] https://laws-lois.justice.gc.ca/eng/

(ECtHR) published in HUDOC,[8] the database of ECtHR.

(k) *Indian Case Law.* We include 34.8K Indian Supreme Court cases originally published by Malik et al. (2021).

The LEXFILES is pre-split into training and test subsets to provide a fair ground for comparing the performance of PLMs that have not been trained in the training set. We use the training subset of the LEXFILES corpus to train 2 new transformer-based languages models, dubbed LEXLMS (Section 4.1), and evaluate their MLM performance across many other already available PLMs (Section 4.2).

3 LEGALLAMA Benchmark

LAnguage Model Analysis (LAMA) (Petroni et al., 2019) is a probing task that is designed to assess specific capabilities of PLMs. The general framework of LAMA is to let PLMs predict a target token behind a `[MASK]` given its context, e.g., *"Paris is the capital of* `[MASK]`*"*, where the answer is 'France'. LEGALLAMA is a new probing benchmark suite inspired by this framework. It includes 8 sub-tasks that aim to assess the acquaintance of legal knowledge that PLMs acquired in the pre-training phase in a *zero-shot fashion*. Such tasks cannot be resolved by laypersons or even law professionals that are not experts in the specific fields of law in many cases.[9] The acquaintance of legal knowledge can be interpreted as some form of primitive understanding of the law, for specific aspects in very controlled (limited) settings -limited legal concepts under a specific jurisdiction-. As Sahlgren and Carlsson (2021) mentioned:

> "Rather than asking whether a language model understands or not, we should ask *to what extent, and in which way*, a model understands."

We further extend the LAMA framework by allowing PLMs to predict multi-token targets. Take for example the "*Drug Trafficking*" offence under the "*Drug-Related*" crimes of the US legislation. Using the RoBERTa tokenizer, this term is split into two tokens, that is "*Drug*" and "*Trafficking*". We replace thus the "drug trafficking" phrase with two `[MASK]` tokens, and then ask the model to predict these tokens simultaneously.

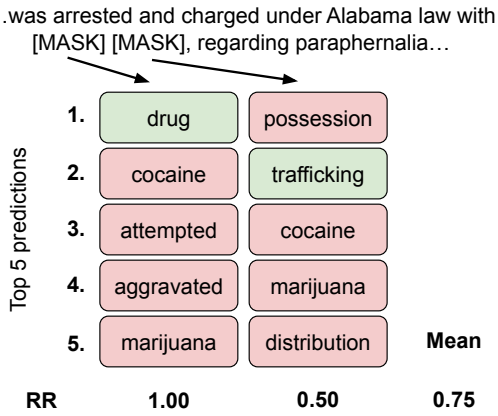

Figure 1: Example from the 'Terminology (US)' subtask. Multi-token LAMA where "drug trafficking" has been replaced with two `[MASK]` tokens. Given the rankings of each predicted token, we compute the reciprocal rank (RR) and obtain a mean reciprocal rank (MRR) over the `[MASK]` tokens.

We evaluate the overall performance of PLMs using the macro-averaged Mean Reciprocal Rank (MRR) (Voorhees and Tice, 2000) over the set of labels (not the entire vocabulary).[10] In the case of multi-token targets, we average the MRR over the predicted tokens.[11] Note that LEGALLAMA examples come from the test subset of the related LexFiles sub-corpora in order to have a fair comparison between models trained or not on the LexFiles training sets. We provide a concrete example in Figure 1, and describe the tasks in detail:

ECHR Articles (CoE). In this task, we have paragraphs from the court assessment section of ECtHR decisions. We extract those paragraphs from the newly introduced ECHR corpus presented in Section 2. The paragraphs include references to ECHR articles, e.g., *"Article* `[MASK]` *of the Convention"*, where `[MASK]` is the article number. For example, *"The applicant complained under Article* ***[2]*** *of the Convention that the prison authorities had failed to protect her son's right to life by taking the necessary measures."* Given a paragraph, where the article number is masked, the model has to predict the associated article number given the context. The dataset is composed of 5,072 test instances containing on average 69 tokens and 13 unique article numbers to predict.

[8] https://hudoc.echr.coe.int/eng

[9] In Appendix A, we present a discussion on the LEGALLAMA tasks' level of difficulty.

[10] We decided to report only MRR results in the main paper for the sake of clarity. Moreover, MRR avoids penalizing for near-identical outcomes. Detailed results including Precision at 1 (P@1) are available in Appendix C.

[11] A stricter evaluation would be to consider a multi-token prediction valid only if all the sub-tokens are properly predicted by the PLM. We decided to average the MRR to consider minor variations and errors.

Contractual Section Titles (US). In this task, we have sections from US contracts reusing the dataset of Tuggener et al. (2020). Contractual sections are usually numbered and titled, e.g., *"10. **[Arbitration]**. Any controversy, dispute or claim directly or indirectly arising out of or relating to this Agreement [...]"*. The section titles reflect the content (subject matter) of the section, and are commonly re-used. Given a section, where the section title is masked, the model has to predict the associated title given the context. The dataset is composed of 1,527 test instances containing on average 85 tokens and 20 unique section titles to predict.

Contract Types (US). In this task, we have introductory paragraphs from US contracts. We extract those paragraphs from the newly introduced corpus of US contracts, presented in Section 2. Introductory paragraphs usually start with the contract title revealing the contract type, e.g., *"Service Agreement"*, and follow with the names of the involved parties, and their roles in this agreement. For example, *"This **[Purchase]** Agreement is entered into this 23rd day of January 2020 by and between A (the "Purchaser") and B (the "Seller")."*. Given an introductory paragraph, where the contract type is masked, the model has to predict the associated type given the context. The task is composed of 1,089 test instances containing on average 150 tokens and 15 unique types of contracts to predict.

Crime Charges (US). In this task, we have paragraphs from US court judgments (opinions). We extract those paragraphs from the US case law corpus, presented in Section 2. We select a list of criminal offenses (e.g., "Sexual Assault"), categorized into 11 major categories (e.g., Sex-related) from the FindLaw website.[12] We filter out paragraphs that refer the specified criminal charges verbatim. For example, *"A person commits the crime of **[burglary]** in the first degree when he or she enters or remains unlawfully in a building with the intent to commit a crime against a person or property therein"* Given a paragraph, where a criminal charge is masked, the model has to predict the associated criminal charge given the context. The task is composed of 4,518 test instances containing on average 118 tokens and 59 charges to predict.

Legal Terminology (US). In this task, we have paragraphs from US court judgments (opinions). We extract those paragraphs from the US case law corpus, presented in Section 2. We select a subset of legal terms per legal topic (e.g., finance law, property law, family law) using the legal vocabularies provided by the Legal Information Institute (LII) of the Cornell Law School.[13] We filter out paragraphs that use the specified legal terms. For example, *"The **[marital privilege]** against self-incrimination is [...] grounded upon the theory that just as one may not be convicted by his own compelled testimony, so may he not be convicted by the testimony of his spouse."* Given a paragraph, where a legal term is masked, the model has to predict the associated legal term given the context. The task is composed of 5,829 test instances containing on average 308 tokens and 92 legal terms from 7 topics to predict.

Legal Terminology (EU). In this task, we have paragraphs from CJEU judgments (opinions). We extract those paragraphs from the newly introduced EU case law corpus, presented in Section 2. We select a subset of legal terms based on the subject matters provided by the database of the courts (CURIA).[14] We filter out paragraphs that use the specified legal terms. For example, *"The guiding principle at the basis of EU **[data protection]** law is that of a self-determined decision of an individual who is capable of making choices about the use and processing of his or her data."* Given a paragraph, where a legal term is masked, the model has to predict the associated legal term given the context. The task is composed of 2,127 test instances containing on average 164 tokens and 42 legal terms from 23 topics to predict.

Legal Terminology (CoE). In this task, we have paragraphs from ECtHR decisions. We extract those paragraphs from the newly introduced ECHR corpus presented in Section 2. We select a subset of legal terms (legal issues) based on the keywords provided by the database of the courts (HUDOC).[15] We filter out paragraphs that use the specified legal terms. For example, *"The applicants alleged that their relatives' **[right to life]** was violated in that they were deliberately killed by village guards."* Given a paragraph, where a legal term is masked, the model has to predict the associated legal term given the context. The task is composed of 6,803

[12] https://www.findlaw.com/criminal/criminal-charges.html

[13] https://www.law.cornell.edu/

[14] https://curia.europa.eu/

[15] https://www.echr.coe.int/Documents/HUDOC_Keywords_ENG.pdf

Model (Source)		# Params	# Vocab	# Acc. Tokens	Pre-training Corpora	
RoBERTa	(Liu et al., 2019)	124/355M	50K	2T	(160GB)	Generic Corpora
LegalBERT	(Chalkidis et al., 2020)	110M	32K	43B	(12GB)	Legal Corpora
CaseLawBERT	(Zheng et al., 2021)	110M	32K	43B	(37GB)	US Case Law
PoL-BERT	(Henderson* et al., 2022)	340M	32K	130B	(256GB)	US Legal Corpora
LexLM	(ours)	124/355M	50K	2T + 256B	(175GB)	Legal Corpora

Table 2: Key specifications of the examined models. We report the number of parameters, the size of vocabulary, the number of accumulated training tokens, and the nature of pre-trainig corpora.

test instances containing on average 97 tokens and 250 legal terms from 15 articles to predict.

Criminal Code Sections (Canada). In this task, we have paragraphs from the Criminal Court of Canada's decisions containing Section Numbers of the Criminal Code of Canada (CCC)[16]. For example, *"Section **[680]** of the Criminal Code provides that a bail review is to be conducted by a panel of this court where directed by the Chief Justice."* Given a paragraph, where a criminal code's section is masked, the model has to predict the associated section number, paragraph, and sub-paragraph (if any) given the context. The task is composed of 321 test instances containing on average 72 tokens and 144 different section numbers to predict.

In Appendix D, we present the full list of vocabulary (masked terms) grouped in categories (clusters) -when applicable- per LegalLAMA sub-task.

4 Experiments

4.1 Pre-trained Language Models

We consider 7 large language models to assess their performance with respect to the upstream (MLM), probing, and downstream evaluation:

RoBERTa (Base/Large) are the original RoBERTa models (Liu et al., 2019) trained for 64k steps with very large batches on generic corpora; thus do not have any clear legal prior (knowledge).

LegalBERT (Base) is a legal-oriented BERT model (Devlin et al., 2019) released by Chalkidis et al. (2020) trained for 1M steps on legal corpora from EU, UK, CoE, and USA.

CaseLawBERT (Base) is another legal-oriented BERT released by Zheng et al. (2021). CaseLawBERT (which we will refer to as *CL-BERT* henceforth) is trained from scratch for 2M steps on the Harvard Law case corpus, which comprises 3.4M legal decisions from US federal and state courts.

[16] https://laws-lois.justice.gc.ca/eng/acts/c-46/index.html

PoL-BERT (Large) is a legal-oriented RoBERTa model released by Henderson* et al. (2022) trained from scratch for 2M steps on the Pile of Law, a corpus consisting of approx. 256GB of English, mainly US, language legal and administrative text.

LexLM (Base/Large) are our newly released RoBERTa models. We follow a series of best-practices in language model development:

(a) We warm-start (initialize) our models from the original RoBERTa checkpoints (base or large) of Liu et al. (2019).

(b) We train a new tokenizer of 50k BPEs, but we reuse the original embeddings for all lexically overlapping tokens (Pfeiffer et al., 2021).

(c) We continue pre-training our models on the diverse LeXFiles (Section 2) corpus for additional 1M steps with batches of 512 samples, and a 20/30% masking rate (Wettig et al., 2023), for base/large models, respectively.

(d) We use a sentence sampler with exponential smoothing of the sub-corpora sampling rate following Conneau et al. (2019) since there is a disparate proportion of tokens across sub-corpora (Table 1) and we aim to preserve per-corpus capacity (avoid overfitting).

(e) We consider mixed cased models, similar to all recently developed large PLMs.

Additional details on LexLM models pre-training can be found in Appendix B.

4.2 Upstream Evaluation

In Table 3, we present the upstream (MLM) performance for all PLMs across the LeXFiles sub-corpora. The performance is measured in terms of accuracy, i.e. Precision@1 of the masked token to be predicted. The accuracy is thus averaged over all the masked tokens for each task. We also provide the average across all tasks, per model. We observe that results vary across models trained in very different settings (model's capacity, pre-

Sub-Corpus	RoBERTa-B	RoBERTa-L	LegalBERT	CL-BERT	PoL-BERT	LexLM-B	LexLM-L
EU Legislation	72.0	75.1	**83.1**	61.4	73.3	78.7	81.8
EU Case Law	72.7	76.5	81.4	63.0	68.5	79.8	**82.9**
UK Legislation	71.3	75.1	86.2	65.1	72.8	84.1	**87.3**
UK Case Law	68.9	73.2	72.3	61.2	62.4	73.2	**76.9**
CAN Legislation	75.5	78.9	80.6	66.4	73.3	82.9	**85.2**
CAN Case Law	62.8	66.0	73.8	64.1	66.0	76.7	**80.3**
US Case Law	68.2	72.5	71.6	64.4	63.8	71.7	**74.8**
US Legislation	74.5	78.1	79.7	65.3	77.0	80.5	**83.5**
US Contracts	67.5	70.9	**89.1**	69.5	76.9	85.1	87.8
ECtHR Case Law	72.0	75.7	**83.3**	61.9	66.3	80.1	**83.3**
Indian Case Law	65.6	70.0	65.2	56.3	58.3	73.3	**76.2**
Average	70.1	73.8	78.7	63.5	68.9	78.7	**81.8**
Model Rank	5	4	2	7	6	2	1

Table 3: Upstream evaluation measured in terms of accuracy (Precision@1) on the Masked Language Modelling (MLM) task across all LEXFILES sub-corpora.

training corpora), while the results also vary across legal sub-corpora.

We want to remind the reader that the upstream evaluation offers a rough idea of a model's capabilities since it relies on random masked sub-words, in which case many of those can be generic and thus highly predictable (e.g. preposition "of"). This phenomenon further motivates the construction of the LEGALLAMA benchmark, in which case only "legal knowledge sensitive" words have been masked.

Type of Documents. In terms of differences across sub-corpora, we observe that the performance on legislation is better compared to case law in 3/4 legal systems, where we have both (EU, UK, US, Canada), with US contractual language being the most predictable for the models which have been trained on it (LexLMs, LegalBERT).

Comparison of PLMs. Overall, the large LexLM model outperforms the rest, being 3% more accurate on average compared to the 2nd best models (base versions of LexLM, and LegalBERT). Such results are expected since LexLMs have been trained in a diverse corpus, similarly to LegalBERT, compared to CL-BERT, and PoL-BERT, which have been trained on US corpora. Over-specialization harms the two US-centric models in a great extend since they are outperformed even from the generic RoBERTa models.

We also observe that LegalBERT outperforms the similarly-sized LexLM in specific sub-corpora (Both EU, UK legislation, ECtHR case law, and US Contracts) that were included in its training. We hypothesize that these results are related to the pre-training data diversity, since LexLMs have been trained in a more diverse corpus including many more documents from different legal systems with a sampling smoothing to preserve capacity per sub-corpus. The larger LexLM model has the capacity to cover all sub-corpora to a greater detail.

In general, larger models pre-trained on the same corpora (RoBERTas, LexLMs) perform better compared to smaller ones, but in-domain pre-training is a much more important factor for upstream performance, e.g., LegalBERT outperforms RoBERTa-L.

4.3 Probing Evaluation

In Table 4, we present the results across all examined PLMs on LEGALLAMA. We analyze the results from two core perspectives: the prior knowledge and the probing task.

Prior Knowledge. The pre-training corpus has a significant impact on the probing performance. RoBERTa models, having little to no legal prior, were expected to achieve worst performance on all probing tasks. Surprisingly, CL-BERT and PoL-BERT achieve on-par or sometimes worst performance than RoBERTa (Base & Large) in most tasks. Being trained on the "Harvard Law Case" corpus (CL-BERT) and the PILE OF LAW (PoL-BERT), we would have expected better performance than a model without legal prior. Their pre-training corpora might be lacking diversity, which might cause their poor performance even on Legal-US probing

Task	Statistics			Models						
	#T	#L	#T/L	RoBERTa-B	RoBERTa-L	LegalBERT	CL-BERT	PoL-BERT	LexLM-B	LexLM-L
ECHR Articles	69	13	1.0	39.8	41.3	91.1	37.5	35.2	91.4	**94.3**
Contract Sections	85	20	1.3	23.6	44.5	80.2	29.2	64.8	**88.2**	87.3
Contract Types	150	15	1.1	43.4	47.8	82.2	54.9	49.7	84.0	**86.1**
Crime Charges (US)	118	59	2.1	56.3	62.4	51.5	62.6	43.5	63.0	**68.1**
Terminology (US)	92	7	2.9	47.1	54.2	60.5	66.7	44.6	66.4	**67.5**
Terminology (EU)	164	42	3.0	38.0	45.3	63.2	38.6	36.9	63.1	**70.4**
Terminology (CoE)	97	250	1.2	45.4	53.1	77.3	49.7	32.8	81.3	**86.8**
CC Sections	72	144	2.0	15.8	19.7	21.9	18.4	19.9	50.6	**68.8**
			Average	33.1	41.3	54.8	38.0	36.8	70.8	**77.4**
			Model Rank	7	4	3	5	6	2	1

Table 4: The 8 LEGALLAMA tasks' statistics regarding the average number of tokens in the input (#T), the number of labels to predict from (#L), and the average number of tokens per label (#T/L) along with the Mean Reciprocal Rank results of the 7 examined PLMs.

tasks. LegalBERT (Base), being trained on UK, EU and USA data illustrates important improvement over models without legal prior (RoBERTa) or having only US legal prior (CaseLaw and PoL-BERT). LEXLM models, being trained on the new LEXFILES dataset, show performance improvement over LegalBERT across all tasks, especially on the task of predicting Section Numbers of the Criminal Code of Canada. Regarding the size of the model, we are able to compare the cased versions of RoBERTa Base/Large and LexLM Base/Large. As expected, the larger versions offer better performance than the smaller ones on every task.

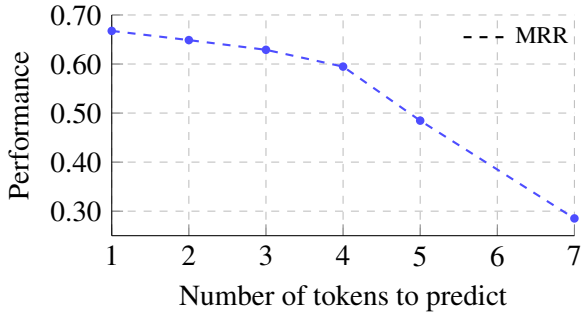

Figure 2: Models performance on LEGALLAMA's test set with respect to the label complexity. Labels with more than three tokens are much harder to predict.

Probing Tasks. We characterize the difficulty of the tasks by their semantic level, the output space (the number of labels to predict from), and the label complexity (how many tokens per label). We expose the tasks' different characteristics in Table 4. Given the best-performing model (LexLM-L), we can see that Crime Charges and Legal Terminology (US and EU) are the hardest tasks to solve. Looking at Table 4, we can see that these three tasks are characterized by a higher label complexity (>2). We further demonstrate the label complexity impact in Figure 2. The output space does not seem to have a correlation with the models' performance, since the selected Legal Terminology Topic Clusters (US) has only 7 possible labels, whereas the Criminal Code Section (Canada) has 144 possible labels. Finally, Crime Charges, being the hardest task to solve, has on average 118 tokens as input and 59 possible labels with moderate complexity, similar to the Terminology tasks (EU and CoE). This suggests that the difficulty of the task is not only driven by the labels' complexity but may rather lie in the lack of contextualization. Take for example the following sentence:

> "This case involves perhaps the first prosecution under New York's new **[computer crime]** statute, Penal Law article 156, which went into effect on November 1, 1986, just days before the incidents charged herein."

The only contextual hint the PLMs have to predict the correct tokens (**[computer crime]**) is the utterance "Penal Law article 156, which went into effect on November 1, 1986". This is the opposite task of predicting article numbers given a context, which is much more difficult than predicting the actual context because the output space is larger.[17]

4.4 Downstream Evaluation

For downstream evaluation, we conduct experiments for 6 legal classification tasks, 5 part of LEXGLUE (Chalkidis et al., 2022a), covering US contracts, US, EU, and ECHR law.

ECtHR (Task B) (Chalkidis et al., 2021b) is a multi-label topic classification task, where given

[17] The actual tokens predicted by the best-performing examined PLM were "sexual" and "abuse".

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Task	μF_1	mF_1	μF_1	mF_1	μF_1	mF_1	μF_1	mF_1	μF_1	mF_1	μF_1	mF_1	μF_1	mF_1
ECtHR	61.2	40.5	74.2	51.5	59.1	37.2	53.6	29.1	69.1	46.9	63.2	41.8	**76.7**	**57.9**
LEDGAR	80.5	62.6	83.6	71.5	81.2	64.7	80.9	64.0	83.3	71.4	82.5	66.8	**84.7**	**72.8**
CNLI	66.8	48.6	68.0	63.5	**70.2**	**65.6**	69.0	64.6	68.3	64.1	61.6	42.9	69.7	64.5
SCOTUS	65.0	36.0	68.9	41.4	60.9	31.2	62.9	33.8	66.3	39.5	66.9	37.7	**71.1**	**43.9**
CaseHOLD	72.7	72.7	75.6	75.6	76.1	76.1	77.6	77.6	73.7	73.7	74.8	74.8	**78.5**	**78.5**
EURLEX	33.4	06.1	62.7	27.1	27.7	04.0	27.0	04.7	60.5	25.4	34.2	06.9	**63.1**	**28.0**
Average	58.4	22.5	71.5	48.6	55.0	17.1	53.9	18.7	69.5	46.4	59.0	24.3	**73.3**	**51.0**
Upstream	5		4		2		7		6		2		1	
Probing	7		4		3		5		6		2		1	
Downstream	5		2		6		7		3		4		1	

Table 5: Test Results for all models across all downstream tasks after fine-tuning for a single epoch.

the facts of an ECtHR case, the model has to predict the alleged violated ECHR article among 10 such articles (e.g., "Art 3. - Prohibition of Torture", "Art. 6 - Right to Fair Trial").

LEDGAR (Tuggener et al., 2020) is a single-label multi-class topic classification task, where given a contractual paragraph, the model has to predict one of the correct topic among 100 topics (e.g., "Limitation of Liability", "Arbitration").

ContractNLI (Koreeda and Manning, 2021) is a contract-based Natural Language Inference (NLI) task, where given an Non-Disclosure Agreement (NDA) and one out 17 templated *hypotheses* (e.g., "The Party may share some Confidential Information with some third-parties."), the model has to predict if the hypothesis is (*entailed*, *contradicted*, or is *neutral*) to the terms of the NDA.

SCOTUS (Chalkidis et al., 2022a) is a single-label multi-class topic classification task, where given a Supreme Court of US (SCOTUS) opinion, the model has to predict the relevant area among 14 issue areas (e.g., "Civil Rights", "Judicial Power").

CaseHOLD (Zheng et al., 2021) is a multiple choice QA classification task, where given a paragraph from a US legal opinion where a legal rule (holding) is masked, the model has to predict the applicable rule among 5 alternatives (the correct one and 2 irrelevant presented in other cases).

EURLEX (Chalkidis et al., 2021a) is a multi-label topic classification task, where given an EU law, the model has to predict the correct EUROVOC concept among hundred concepts (e.g., "Environmental Policy", "International Trade").

We fine-tune all examined PLMs (Section 4.1) for a single epoch with a learning rate of $1e-5$ leading to a small number of updates. We are interested to examine how fast each model convergence based on its prior knowledge; in other words, what can a model learn in a single pass over training data? Finetuning models for many epochs over large datasets will eventually lead to a full re-parameterization of the models, in which case the importance of prior knowledge will diminish compromise the goal of our study (Figure 3).[18]

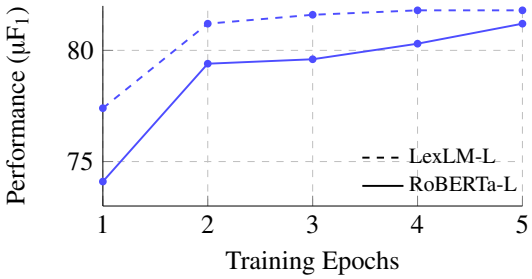

Figure 3: Development Results of RoBERTa and LexLM large on ECtHR across 5 training epochs.

For all tasks, we use standard N-way classifiers with a classification head (Devlin et al., 2019). For ECtHR, and SCOTUS, involving long documents, we warm-start Longformer (Beltagy et al., 2020) models from each PLM's parameters to encode up to 2048 tokens. We evaluate classification performance with micro-F1 (μF_1) and macro-F1 (mF_1) across tasks following Chalkidis et al. (2022a).

Results In Table 5, we present the test results across all tasks/datasets. We analyze the results from two perspectives: model's capacity (size), and prior legal knowledge abducted via pre-training.

[18] In most tasks, models fully converge after approx. 5 epochs with improved performance, and the relative differences between generic and legal-oriented models are diminished (Chalkidis et al., 2022a).

Model's capacity (size) strongly correlates with the overall downstream performance. Across all tasks, there are 2/6 exceptions (CNLI and CaseHOLD) where LegalBERT outperforms larger PLMs. Both tasks are using sentence pairs, a setup used in BERT's pre-training, but not in RoBERTa, which may bring LegalBERT, a BERT-based model, in a better initial condition co-considering the minimal updates steps, compared to all large models following the RoBERTa pre-training setup, which do no use pairs of sentences or optimized based on a sentence-level objective (NSP).

Legal Knowledge also plays an important role following the model's capacity (size). We observe that LexLM-B trained in the diverse LeXFiles corpus outperforms the equally-sized RoBERTa-B model in 5/6 tasks, while LegalBERT and CL-BERT outperform it only in 3 out of 6 tasks. In this case, the results are mixed, i.e., acquaintance of legal knowledge as expressed by upstream (Section 4.2) and probing (Section 4.3) performance does not correlate with downstream performance.

In the case of large-sized models, LexLM-L outperform RoBERTa-L across all tasks, while PoL-BERT trained on the US-biased Pile of Law corpus is outperformed by RoBERTa-L in 5 out of 6 tasks. Given the results with respect to upstream and probing performance, RoBERTa-L has a better legal prior; so in these regards, acquaintance of legal knowledge fully correlates with downstream performance in the large models' regime.

5 Release of Resources

We release our code base to assure reproducibility and let others extend our study by experimenting with other PLMs, or develop new ones.[19] The new LexLM models (Section 4.1), the LeXFiles corpus [20] (Section 2), and the LegalLAMA benchmark [21] (Section 4.3) are available on Hugging Face Hub (Lhoest et al., 2021).[22]

6 Conclusions and Future Work

In this work, we introduced a multinational English legal corpus (LeXFiles) and a legal knowledge probing benchmark (LegalLAMA) to facilitate training and detailed analysis of legal-oriented PLMs. We also released two new legal PLMs and evaluate them alongside others on LegalLAMA and LexGLUE. Based on our analysis (Section 4), we make the following general observations:

(a) The use of diverse legal corpora leads to better overall upstream performance (Section 4.2).
(b) We find that probing performance strongly correlates with upstream performance in related legal topics (Section 4.3).
(c) For both upstream, and probing performance, the selection of pre-training corpora has a much larger effect compared to model's capacity (Sections 4.2-4.3). Nonetheless, larger models pre-trained on similar corpora have better overall performance.
(d) Downstream performance is mainly driven by the model's capacity and prior legal knowledge which can be estimated by upstream and probing performance (Section 4.4).

In future work, we plan to further analyze the learning dynamics of legal language models by comparing their representations with representations derived from legal knowledge bases. Given the availability of the new resources, the development of instruction-following (Wei et al., 2021) fine-tuned legal-oriented GPT-like (Ouyang et al., 2022) models is also an anticipated direction.

Limitations

Diversity of Corpora While the newly introduced LeXFiles corpus is significantly more diverse compared to the Pile of Law corpus of Henderson* et al. (2022), it is still an English-only corpus covering only 6 legal systems (EU, UK, CoE, US, India, Canada). Despite, the fact that we can train better models (LexLMs) and evaluate these models across these corpora, in future work, we should extend our analysis to cover even more languages and legal systems, and a higher granularity in the labeling of legal fields within these systems. Not only will this help support the inclusion of other legal traditions but also adding more linguistic and cultural diversity will help us better understand the robustness of existing methods.

Similarly, the newly introduced LegalLAMA benchmark consists of 8 sub-tasks targeting EU, ECHR, US, and Canadian jurisdictions in a very controlled setting; where examples were automatically extracted. While on this benchmark, legal-oriented PLMs has demonstrated a significant degree of "understanding" of legal language and legal

[19] https://github.com/coastalcph/lexlms
[20] https://huggingface.co/datasets/lexlms/lex_files
[21] https://huggingface.co/datasets/lexlms/legal_lama
[22] https://huggingface.co/lexlms

topics, this benchmark should be further expanded with more sub-tasks to evaluate the acquaintance of legal knowledge across more legal systems and topics, and possibly cleansed from both very easy and unsolvable examples.

Model Considerations In this work, we consider encoder-only (BERT-like) models up to approx. 350M parameters, while recent work on the development of Large Language Models (LLMs) (Kaplan et al., 2020; Brown et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2022) is mainly targeting billion-parameter-sized models (10-100Bs of parameters) that usually follow a decoder-only, e.g., GPT (Radford and Narasimhan, 2018), or encoder-decoder, e.g., T5 (Raffel et al., 2020), architecture. Moreover, new paradigms of training PLMs have been introduced, such as *instruction-based finetuning* (Wei et al., 2021), and *alignment* via Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022). Latest GPT models (Ouyang et al., 2022) have recently shown significant zero-shot progress on law-related tasks such as bar examination question answering (Katz et al., 2023). Thus, future work should follow the most recent advances by pre-training much larger auto-regressive GTP-like models that seem to lead to emergent zero-shot and few-shot capabilities.

Evaluation Considerations In Section 3, we present how we account for and evaluate multi-token expressions (terms) on the LegalLAMA benchmark; we are open to ideas on how we should possibly improve the current approach to provide a fairer and more robust evaluation framework across all models. Similarly, in Section 4.4, we fine-tune all examined PLMs for a single epoch to avoid extreme over-reparameterization and better estimate how model's knowledge affects convergence and performance. Nonetheless, there are possibly better approaches to control for these aspects, e.g., Adapter-based (Rücklé et al., 2021) finetuning, or other approaches, such as LoRA (Hu et al., 2022).

Beyond Performance While we consider a multi-facet analysis, we do not cover other interesting dimensions that should also be explored, especially since law is a very sensitive application domain; for instance trustworthiness-related topics, such as model interpretability (Chalkidis et al., 2021b; Malik et al., 2021), and fairness (Chalkidis et al., 2022b). Future work can build from the results reported herein to explore these important topics.

Ethics Statement

The scope of this work is to examine the performance of legal-oriented PLMs from a multi-facet perspective and broaden the discussion to help practitioners build assisting technology for legal professionals and laypersons. We believe that this is an important application field, where research should be conducted (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize law, while also highlighting (informing the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of legal-oriented technologies.

In this direction, we introduce new resources covering various legal systems to build new models that better represent law and better assess their capabilities. All newly developed and published resources are based on publicly available data, most of them scattered on several web portals.

Acknowledgments

This work was partly funded by the Innovation Fund Denmark (IFD, https://innovationsfonden.dk/en) and the Fonds de recherche du Québec – Nature et technologies (FRQNT, https://frq.gouv.qc.ca/nature-et-technologies/).

References

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

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

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset

for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021b. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Lawrence M. Friedman and Grant M. Hayden. 2017. 1What Is a Legal System? In *American Law: An Introduction*. Oxford University Press.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2022. Mining Legal Arguments in Court Decisions. *arXiv preprint*.

Peter Henderson*, Mark S. Krass*, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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 James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. 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.

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*.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain.

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Magnus Sahlgren and Fredrik Carlsson. 2021. The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Ellen Voorhees and D Tice. 2000. The trec-8 question answering track evaluation. 3. The TREC-8 Question Answering Track Evaluation.

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

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of*

the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

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.

A LegalLAMA Discussion

The LEGALLAMA tasks cannot be resolved by laypersons or even law professionals that are not experts in the specific fields of law in many cases. Another consideration that often goes unspecified is that expertise is legal system-specific (e.g. US law differs widely from EU law), as do the distinctions between the academic and the practical knowledge of law (including potential sub-distinctions between different types of legal practitioners, e.g. litigation experts, contract drafting experts, due diligence experts, etc.). Lastly, it is also important to note that legal systems can be clustered according to similarities or differences. Specifically:

- For task **'ECHR Articles'**, both laypersons and lawyers who are not experts in human rights law (particularly ECHR) would perform at random chance level, since they lack knowledge of the ECHR in an article level. Providing the titles of the articles (Table 6), we can expect improved performance in case of rich context. Generally, the same can be said for the related task **'Legal Terminology (CoE)'**. Legal terminology is very particular to individual legal systems, and predicting the place of legal concepts within the ECHR would require a very high level of specialization.
- For task **'Contractual Section Titles (US)'**, structural knowledge of US contracts would be necessary for the performance of this task with a high degree of accuracy. This is due to the fact that contracts often have some structural similarities, but also particular characteristics depending on the type of contract (e.g. employment, sale, credit). Laypersons would perform this task at random chance level. Practicing lawyers with contract drafting expertise would potentially have the highest performance in this task. Non-US lawyers with no contract drafting expertise would perform slightly higher than random chance level. The same considerations apply to the task **'Contract Types (US)'**.
- For tasks **'Crime Charges (US)'** and **'Criminal Code Sections (Canada)'**, both laypersons and lawyers who are not experts in criminal law (particularly US law and Canadian law) would perform at random chance level, since the legal concepts are very specific (e.g. manslaughter). Improved performance could be seen in cases where the masked terms are specifically defined.
- For tasks **'Legal Terminology (US)'** and **'Legal Terminology (EU)'**, the same discussion as above is applicable. Legal terminology is system-specific. There may be similar terms, but in the absence of knowledge relating to how such similarities may be interpreted, a non-expert lawyer would not perform such a task with a very high accuracy level.

A.1 ECtHR Articles

We hereby provide details on the 13 ECtHR articles;

ECHR Article	Description (Title)
Article 2	Right to life
Article 3	Prohibition of torture
Article 5	Right to liberty and security
Article 6	Right to a fair trial
Article 7	No punishment without law
Article 8	Right to respect for private and family life
Article 9	Freedom of thought, conscience and religion
Article 10	Freedom of expression
Article 11	Freedom of assembly and association
Article 13	Right to an effective remedy
Article 14	Prohibition of discrimination
Article 34	Individual applications
Article 35	Admissibility criteria

Table 6: ECHR Articles

B LexLM Pre-training Details

For the newly released, LexLM models (LexLMs), we followed a series of best-practices in language model development literature:

(a) We warm-start (initialize) our models from the original RoBERTa checkpoints (base or large) of Liu et al. (2019). Model recycling

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Task	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
ECHR Articles	0.26	0.40	0.27	0.41	0.86	0.91	0.23	0.38	0.20	0.35	0.86	0.91	**0.91**	**0.94**
Contract Sections	0.20	0.40	0.53	0.66	0.77	0.85	0.24	0.40	0.51	0.65	**0.78**	**0.86**	**0.78**	**0.86**
Contract Types	0.32	0.48	0.34	0.50	0.80	0.87	0.42	0.55	0.37	0.50	0.82	0.89	**0.85**	**0.91**
Crime Charges (US)	0.46	0.58	0.54	0.65	0.44	0.56	0.51	0.63	0.33	0.45	0.56	0.67	**0.61**	**0.71**
Terminology (US)	0.41	0.51	0.49	0.58	0.52	0.63	0.58	0.69	0.37	0.49	0.64	0.74	**0.70**	**0.79**
Terminology (EU)	0.34	0.47	0.40	0.53	0.51	0.64	0.25	0.39	0.25	0.38	0.60	0.72	**0.67**	**0.77**
Terminology (CoE)	0.43	0.54	0.51	0.60	0.69	0.78	0.36	0.49	0.30	0.41	0.78	0.86	**0.86**	**0.91**
CC Sections	0.36	0.45	0.40	0.50	0.53	0.59	0.45	0.54	0.46	0.53	0.77	0.83	**0.86**	**0.90**
Average	0.33	0.47	0.41	0.54	0.61	0.71	0.34	0.49	0.32	0.46	0.71	0.80	**0.77**	**0.85**
Model Rank	6		4		3		5		7		2		1	

Table 7: P@1 and MRR results of the 7 examined PLMs on the 8 LEGALLAMA tasks.

is a standard process followed by many (Wei et al., 2021; Ouyang et al., 2022) to benefit from starting from an available "well-trained" PLM, instead from scratch (random).

(b) We train a new tokenizer of 50k BPEs based on the training subsets of LEXFILES to better cover legal language across all covered legal systems. Although, we reuse the original RoBERTa embeddings for all lexically overlapping tokens (Pfeiffer et al., 2021), i.e., we warm-start word embeddings for tokens that already exist in the original RoBERTa vocabulary, and use random ones for the rest.

(c) We continue pre-training our models on the diverse LEXFILES (Section 2) corpus for additional 1M steps with batches of 512 samples. We do initial warm-up steps for the first 5% of the total training steps with a linearly increasing learning rate up to $1e-4$, and then follow a cosine decay scheduling, following recent trends. For half of the warm-up phase (2.5%), the Transformer encoder is frozen, and only the embeddings, shared between input and output (MLM), are updated. We also use an increased 20/30% masking rate, where also 100% of the predictions are based on masked tokens, compared to Devlin et al. (2019)[23] for base/large models respectively, based on the findings of Wettig et al. (2023).

(d) For both training the tokenizer and the LexLM models, we use a sentence sampler with exponential smoothing of the sub-corpora sampling rate following Conneau et al. (2019) and Raffel et al. (2020), since there is a disparate proportion of tokens across sub-corpora (Table 1) and we aim to preserve per-corpus capacity, i.e., avoid overfitting to the majority (approx. 94% of the total number or tokens) US-origin texts.

(e) We consider mixed cased models, similar to all recently developed large PLMs (Liu et al., 2019; Raffel et al., 2020; Brown et al., 2020).

We make LexLM models (base/large) publicly available alongside all intermediate checkpoints every 50k training steps on Hugging Face Hub.[24]

C Detailed Legal-LAMA results per tasks

Table 7 contains the same results as in Table 4 with the addition of Precision@1 scores (P@1). The reason why we decided to only present MRR results in the main paper is that the difference between MRR and P@1 does not change the ranking of the models, and P@1 does not account for minor variations in predictions.

For each task, we display detailed results per predicted terms for each model. Table 8 contains results on the 13 article numbers from the ECHR task. Table 9 contains results on the 20 clause types from the Contract Section task. Table 10 contains results on the 16 types of contracts from the Contract Section task. Table 11 contains results on the 11 topics from the Crime Charges (US) task. Each topic contains multiple labels. Table 12 contains results on the 7 topics from the Terminology (US) task. Each topic contains multiple labels. Table 13

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

[24] https://huggingface.co/lexlms

contains results on the 23 topics from the Terminology (EU) task. Each topic contains multiple labels. Table 14 contains results on the 12 articles from the Terminology (CoE) task. Each article contains multiple labels. Table 15 contains results on the 43 sections from the Criminal Code Sections (Canada) task.

D LegalLAMA Tasks' Vocabulary

In Tables 8, 9, 10, 13, and 15 we present the labels' list for the 'ECHR Articles', 'Contract Sections', 'Contract Types', 'Terminology (EU)' and 'Criminal Code Sections (Canada)' sub-tasks and the label-wise performance. In Tables 16, 17, and 18, we present the labels' list for the 'Terminology (CoE)', 'Crimes Charges (US)', and 'Terminology (US)' sub-tasks grouped in clusters.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
ECHR Article	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
Art. 2	0.87	0.91	0.63	0.76	0.87	0.92	0.27	0.45	0.29	0.51	0.86	0.91	0.91	0.94
Art. 3	0.23	0.56	0.35	0.59	0.93	0.96	0.44	0.62	0.32	0.54	0.93	0.96	0.96	0.97
Art. 5	0.35	0.56	0.39	0.58	0.83	0.89	0.32	0.44	0.20	0.41	0.79	0.86	0.88	0.92
Art. 6	0.27	0.40	0.26	0.38	0.93	0.96	0.28	0.43	0.18	0.36	0.93	0.96	0.94	0.96
Art. 7	0.15	0.38	0.30	0.53	0.53	0.72	0.15	0.36	0.29	0.49	0.62	0.75	0.74	0.83
Art. 8	0.16	0.28	0.18	0.36	0.89	0.93	0.17	0.32	0.13	0.30	0.89	0.94	0.91	0.95
Art. 9	0.33	0.46	0.32	0.46	0.83	0.89	0.27	0.45	0.27	0.45	0.85	0.92	0.95	0.97
Art. 10	0.23	0.34	0.24	0.37	0.84	0.90	0.27	0.43	0.21	0.33	0.87	0.91	0.90	0.93
Art. 11	0.25	0.33	0.27	0.36	0.94	0.96	0.30	0.44	0.23	0.34	0.91	0.94	0.97	0.99
Art. 13	0.28	0.36	0.32	0.40	0.89	0.94	0.27	0.36	0.26	0.39	0.90	0.94	0.92	0.95
Art. 14	0.14	0.24	0.15	0.26	0.85	0.91	0.14	0.27	0.07	0.19	0.88	0.92	0.90	0.94
Art. 34	0.09	0.20	0.08	0.19	0.90	0.93	0.08	0.17	0.06	0.15	0.90	0.94	0.93	0.96
Art. 35	0.05	0.13	0.06	0.17	0.90	0.94	0.05	0.13	0.05	0.13	0.88	0.93	0.92	0.95
Average	0.26	0.40	0.27	0.41	0.86	0.91	0.23	0.38	0.20	0.35	0.86	0.91	0.91	0.94

Table 8: P@1 and MRR results of the 7 examined PLMs on the 13 article numbers from the ECHR task.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Clause Type	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
Arbitration	0.44	0.65	0.97	0.98	1.00	1.00	0.83	0.91	1.00	1.00	1.00	1.00	1.00	1.00
Assignments	0.05	0.15	0.34	0.49	0.85	0.89	0.01	0.12	0.40	0.58	0.90	0.94	0.94	0.96
Confidentiality	0.14	0.34	0.73	0.84	0.99	0.99	0.14	0.34	0.67	0.77	0.99	0.99	0.99	0.99
Costs	0.00	0.22	0.56	0.66	0.78	0.89	0.22	0.38	0.33	0.54	0.56	0.78	0.67	0.80
Definitions	1.00	1.00	0.99	0.99	0.78	0.84	0.27	0.53	0.75	0.85	0.78	0.85	0.81	0.87
Disclosures	0.56	0.70	0.37	0.50	0.80	0.89	0.02	0.16	0.01	0.23	0.65	0.80	0.59	0.77
Employment	0.42	0.69	1.00	1.00	0.92	0.96	0.50	0.67	0.65	0.80	0.85	0.92	1.00	1.00
Enforceability	0.00	0.17	0.26	0.37	0.42	0.64	0.00	0.06	0.25	0.42	0.33	0.54	0.16	0.39
Fees	0.12	0.50	0.52	0.70	0.43	0.62	0.39	0.54	0.38	0.60	0.48	0.67	0.51	0.69
Indemnification	0.41	0.59	0.70	0.80	0.92	0.96	0.10	0.34	0.98	0.98	0.96	0.98	0.97	0.98
Law	0.00	0.40	0.21	0.57	0.37	0.58	0.87	0.92	0.00	0.16	0.79	0.87	0.78	0.86
Participations	0.04	0.20	0.45	0.66	0.82	0.90	0.52	0.67	0.38	0.59	0.80	0.87	0.82	0.89
Remedies	0.05	0.25	0.16	0.34	0.92	0.96	0.11	0.37	0.52	0.71	0.98	0.99	0.99	0.99
Representations	0.01	0.30	0.43	0.62	0.77	0.85	0.17	0.46	0.46	0.64	0.86	0.91	0.80	0.87
Severability	0.02	0.17	0.34	0.58	0.99	0.99	0.00	0.16	0.97	0.98	0.98	0.99	0.98	0.99
Solvency	0.09	0.22	0.38	0.52	0.94	0.97	0.00	0.06	0.11	0.26	0.97	0.99	0.97	0.99
Taxes	0.29	0.59	0.86	0.90	0.99	0.99	0.24	0.48	0.56	0.68	0.99	0.99	0.99	0.99
Termination	0.31	0.56	0.60	0.77	0.75	0.85	0.22	0.45	0.84	0.91	0.80	0.89	0.76	0.86
Waivers	0.12	0.22	0.59	0.67	0.79	0.87	0.00	0.07	0.57	0.74	0.94	0.95	0.84	0.89
Warranties	0.00	0.14	0.05	0.26	0.08	0.39	0.14	0.33	0.27	0.53	0.05	0.36	0.10	0.41
Average	0.20	40.2	0.53	0.66	0.77	0.85	0.24	0.40	0.51	0.65	0.78	0.86	0.78	0.86

Table 9: P@1 and MRR results of the 7 examined PLMs on the 20 clause types from the Contract Section task.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Contract Type	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
Award	0.62	0.67	0.62	0.70	1.00	1.00	0.54	0.60	0.62	0.70	1.00	1.00	1.00	1.00
Consulting	0.03	0.17	0.10	0.23	0.94	0.97	0.08	0.29	0.07	0.17	0.81	0.87	0.90	0.93
Credit	0.57	0.72	0.37	0.53	0.97	0.98	0.80	0.88	0.55	0.77	0.90	0.95	0.95	0.98
Employment	0.40	0.54	0.30	0.44	0.88	0.94	0.63	0.73	0.56	0.72	0.99	0.99	0.96	0.98
Indemnity	0.08	0.34	0.00	0.16	0.62	0.71	0.00	0.15	0.00	0.11	1.00	1.00	1.00	1.00
Letter	0.22	0.33	0.24	0.34	0.96	0.98	0.76	0.87	0.18	0.27	0.77	0.88	0.93	0.97
License	0.40	0.62	0.20	0.42	0.63	0.76	0.49	0.70	0.31	0.44	0.69	0.79	0.86	0.91
Loan	0.51	0.67	0.72	0.84	0.90	0.93	0.72	0.83	0.95	0.97	0.90	0.94	0.87	0.93
Purchase	0.70	0.83	0.59	0.68	0.70	0.83	0.52	0.68	0.93	0.96	0.89	0.92	0.93	0.94
Security	0.35	0.56	0.70	0.80	0.95	0.97	0.59	0.75	0.35	0.59	0.97	0.99	0.97	0.99
Separation	0.12	0.26	0.16	0.28	0.66	0.77	0.15	0.38	0.07	0.21	0.73	0.86	0.71	0.82
Services	0.24	0.45	0.29	0.48	0.52	0.67	0.05	0.19	0.38	0.54	0.52	0.69	0.52	0.69
Settlement	0.49	0.63	0.49	0.71	0.70	0.80	0.88	0.93	0.58	0.72	0.53	0.74	0.65	0.80
Supply	0.09	0.24	0.35	0.51	0.61	0.73	0.09	0.19	0.04	0.14	0.70	0.77	0.65	0.74
Voting	0.00	0.13	0.03	0.33	1.00	1.00	0.00	0.10	0.00	0.13	0.83	0.91	0.90	0.95
Average	0.32	0.48	0.34	0.50	0.80	0.87	0.42	0.55	0.37	0.50	0.82	0.89	0.85	0.91

Table 10: P@1 and MRR results of the 7 examined PLMs on the 16 types of contracts from the Contract Types task.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Crime Charges	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
Children	0.69	0.78	0.73	0.82	0.47	0.61	0.67	0.78	0.45	0.60	0.73	0.82	0.77	0.85
Computer	0.36	0.51	0.46	0.62	0.32	0.41	0.42	0.53	0.29	0.40	0.44	0.56	0.51	0.64
Court-related	0.55	0.66	0.57	0.69	0.53	0.65	0.61	0.73	0.44	0.58	0.63	0.74	0.67	0.78
Drug-related	0.40	0.53	0.48	0.60	0.31	0.44	0.35	0.50	0.26	0.38	0.42	0.55	0.46	0.60
Wrongful Life Taking	0.50	0.64	0.59	0.72	0.59	0.71	0.58	0.72	0.31	0.47	0.61	0.74	0.63	0.76
Mens Rea	0.56	0.64	0.62	0.69	0.55	0.65	0.68	0.76	0.47	0.59	0.69	0.77	0.75	0.82
Monetary	0.40	0.51	0.48	0.59	0.52	0.63	0.50	0.63	0.30	0.44	0.53	0.65	0.61	0.72
Pattern of Behavior	0.37	0.50	0.48	0.59	0.41	0.50	0.44	0.57	0.26	0.37	0.52	0.62	0.57	0.68
Property	0.25	0.34	0.36	0.43	0.26	0.36	0.32	0.41	0.14	0.22	0.40	0.46	0.42	0.48
Sex-related	0.55	0.65	0.59	0.70	0.47	0.59	0.54	0.66	0.36	0.48	0.60	0.70	0.66	0.75
Violent	0.46	0.61	0.57	0.70	0.45	0.59	0.54	0.69	0.29	0.45	0.58	0.72	0.65	0.77
Average	0.46	0.58	0.54	0.65	0.44	0.56	0.51	0.63	0.33	0.45	0.56	0.67	0.61	0.71

Table 11: Results on the 'Crime Charges (US)' LEGALLAMA tasks. Results are clustered in Crime Topics.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Topic	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
Business law	0.29	0.38	0.37	0.45	0.48	0.59	0.59	0.70	0.35	0.46	0.59	0.71	0.69	0.79
Criminal law	0.39	0.49	0.46	0.54	0.48	0.58	0.54	0.65	0.32	0.45	0.64	0.73	0.67	0.76
Employment law	0.47	0.60	0.58	0.68	0.47	0.60	0.54	0.67	0.41	0.54	0.55	0.67	0.65	0.76
Family law	0.52	0.61	0.59	0.67	0.49	0.62	0.66	0.77	0.40	0.52	0.75	0.84	0.82	0.88
Immigration	0.48	0.57	0.54	0.62	0.58	0.67	0.55	0.65	0.38	0.48	0.65	0.74	0.72	0.80
Landlord-tenant law	0.37	0.46	0.44	0.52	0.64	0.73	0.69	0.77	0.42	0.52	0.75	0.82	0.80	0.86
Bankruptcy	0.37	0.49	0.43	0.55	0.48	0.59	0.49	0.62	0.34	0.47	0.53	0.66	0.59	0.71
Average	0.41	0.51	0.49	0.58	0.52	0.63	0.58	0.69	0.37	0.49	0.64	0.74	0.70	0.79

Table 12: Results on the 'Terminology (US)' LEGALLAMA task. Results are clustered in Law Topics.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Topic	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
Accession	0.32	0.45	0.57	0.68	0.93	0.95	0.46	0.55	0.87	0.90	0.80	0.88	0.80	0.89
Administrative cooperation	0.15	0.33	0.23	0.40	0.53	0.69	0.12	0.27	0.19	0.32	0.65	0.79	0.82	0.89
Approximation of laws	0.46	0.54	0.54	0.58	0.36	0.47	0.18	0.32	0.08	0.23	0.67	0.73	0.72	0.79
Area of freedom, security and justice	0.14	0.27	0.13	0.28	0.11	0.24	0.14	0.28	0.11	0.25	0.13	0.27	0.19	0.34
Citizenship of the union	0.40	0.60	0.47	0.64	0.26	0.45	0.12	0.30	0.31	0.47	0.50	0.70	0.53	0.72
Competition	0.50	0.68	0.75	0.80	0.84	0.90	0.52	0.62	0.52	0.62	0.88	0.89	0.88	0.89
Consumer protection	0.40	0.57	0.50	0.62	0.45	0.58	0.28	0.42	0.20	0.37	0.25	0.42	0.40	0.54
Data protection	0.47	0.63	0.61	0.73	0.64	0.75	0.17	0.28	0.20	0.35	0.66	0.76	0.73	0.82
External relations	0.30	0.45	0.40	0.61	0.38	0.55	0.19	0.29	0.09	0.22	0.40	0.61	0.55	0.68
Free movement of capital	0.42	0.45	0.42	0.45	0.18	0.38	0.11	0.26	0.08	0.22	0.33	0.53	0.33	0.59
Free movement of goods	0.25	0.37	0.25	0.35	0.32	0.48	0.21	0.34	0.18	0.31	0.62	0.74	0.38	0.58
Freedom of establishment	0.22	0.34	0.42	0.50	0.64	0.75	0.33	0.43	0.29	0.40	0.81	0.88	0.94	0.95
Freedom of movement for workers	0.22	0.34	0.35	0.41	0.19	0.35	0.12	0.23	0.11	0.22	0.43	0.56	0.38	0.55
Freedom to provide services	0.07	0.20	0.04	0.23	0.23	0.40	0.10	0.24	0.15	0.29	0.39	0.58	0.54	0.67
Fundamental rights	0.60	0.73	0.69	0.81	0.89	0.93	0.26	0.37	0.22	0.36	0.84	0.90	0.83	0.89
Internal market	0.00	0.24	0.20	0.40	0.94	0.96	0.26	0.36	0.40	0.55	0.40	0.62	0.70	0.77
Non-contractual liability	0.09	0.19	0.09	0.20	0.19	0.35	0.19	0.40	0.10	0.23	0.30	0.49	0.55	0.70
Non-discrimination	0.00	0.24	0.00	0.25	0.50	0.68	0.29	0.48	0.10	0.26	0.67	0.83	0.33	0.67
Privileges and immunities	0.17	0.27	0.12	0.24	0.63	0.77	0.25	0.36	0.20	0.35	0.81	0.88	0.81	0.87
Procedural provisions	0.53	0.66	0.63	0.75	0.68	0.80	0.61	0.73	0.42	0.56	0.71	0.82	0.75	0.84
Public health	0.62	0.80	0.50	0.72	0.68	0.79	0.38	0.58	0.28	0.48	0.54	0.75	0.92	0.96
Safeguard measures	0.50	0.52	0.50	0.58	0.64	0.76	0.31	0.39	0.42	0.52	0.75	0.88	1.00	1.00
Social policy	0.75	0.78	0.75	0.81	0.42	0.54	0.22	0.37	0.15	0.32	0.75	0.83	1.00	1.00
Average	0.34	0.47	0.40	0.53	0.51	0.64	0.25	0.39	0.25	0.38	0.60	0.72	0.67	0.77

Table 13: Results on the 'Terminology (EU)' LEGALLAMA task. Results are clustered in Law Topics.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Article	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
Art. 2	0.46	0.57	0.52	0.63	0.72	0.82	0.37	0.51	0.36	0.47	0.80	0.87	0.90	0.94
Art. 3	0.51	0.61	0.58	0.69	0.80	0.87	0.40	0.54	0.34	0.45	0.83	0.90	0.89	0.93
Art. 5	0.39	0.51	0.46	0.57	0.56	0.69	0.36	0.48	0.25	0.38	0.63	0.75	0.74	0.83
Art. 6	0.42	0.55	0.49	0.62	0.68	0.77	0.43	0.55	0.36	0.49	0.77	0.85	0.82	0.89
Art. 7	0.71	0.78	0.82	0.86	0.89	0.93	0.36	0.59	0.44	0.52	0.88	0.93	0.91	0.94
Art. 8	0.35	0.47	0.45	0.56	0.62	0.71	0.29	0.41	0.26	0.36	0.73	0.82	0.84	0.90
Art. 9	0.49	0.57	0.56	0.64	0.67	0.76	0.43	0.53	0.33	0.44	0.79	0.86	0.85	0.91
Art. 10	0.30	0.43	0.41	0.52	0.57	0.69	0.25	0.37	0.20	0.31	0.73	0.82	0.84	0.90
Art. 11	0.32	0.44	0.42	0.52	0.66	0.75	0.29	0.40	0.23	0.34	0.74	0.84	0.87	0.92
Art. 13	0.44	0.61	0.55	0.69	0.78	0.86	0.38	0.56	0.27	0.45	0.86	0.90	0.91	0.94
Art. 14	0.72	0.80	0.79	0.85	0.80	0.86	0.69	0.78	0.52	0.63	0.84	0.89	0.91	0.94
Art. 35	0.14	0.21	0.18	0.24	0.61	0.71	0.14	0.26	0.09	0.18	0.78	0.85	0.89	0.93
Average	0.43	0.54	0.51	0.61	0.69	0.79	0.36	0.49	0.30	0.41	0.78	0.86	0.86	0.91

Table 14: Results on the 'Terminology (CoE)' LEGALLAMA task. Results are clustered by Article.

	RoBERTa-B		RoBERTa-L		LegalBERT		CL-BERT		PoL-BERT		LexLM-B		LexLM-L	
Section	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR	P@1	MRR
16	0.00	0.08	0.00	0.04	0.00	0.04	0.00	0.10	0.00	0.08	0.50	0.62	1.00	1.00
21	0.23	0.41	0.37	0.47	0.46	0.56	0.43	0.56	0.44	0.55	0.94	0.96	0.97	0.99
85	0.46	0.51	0.31	0.42	0.30	0.41	0.29	0.37	0.30	0.39	0.40	0.52	0.57	0.69
86	0.38	0.53	0.38	0.50	0.50	0.62	0.50	0.54	0.50	0.53	0.50	0.71	0.50	0.66
87	0.75	0.78	0.50	0.62	0.75	0.79	0.50	0.65	0.75	0.82	0.75	0.83	0.75	0.80
88.23	0.25	0.34	0.33	0.38	0.33	0.38	0.33	0.39	0.33	0.42	0.33	0.40	0.33	0.38
95	0.48	0.54	0.52	0.56	0.52	0.55	0.46	0.52	0.45	0.49	0.79	0.84	0.80	0.85
122	0.17	0.19	0.11	0.15	0.17	0.18	0.12	0.15	0.12	0.16	0.50	0.67	0.83	0.86
145	0.25	0.38	0.25	0.40	0.44	0.51	0.38	0.50	0.50	0.55	0.62	0.71	0.88	0.90
151	0.59	0.61	0.89	0.91	0.62	0.64	0.04	0.34	0.02	0.32	0.91	0.92	0.91	0.92
152	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.75	1.00	1.00	1.00	1.00
163	0.50	0.52	0.50	0.51	0.50	0.51	0.50	0.51	0.50	0.52	0.50	0.75	1.00	1.00
163.1	0.33	0.40	0.44	0.57	0.67	0.68	0.33	0.51	0.33	0.46	1.00	1.00	1.00	1.00
231	0.25	0.29	0.38	0.54	0.62	0.65	0.44	0.51	0.56	0.59	0.94	0.94	1.00	1.00
249	0.40	0.45	0.33	0.41	0.60	0.68	0.66	0.74	0.53	0.66	0.87	0.91	0.88	0.90
254	0.50	0.61	0.65	0.73	0.50	0.58	0.40	0.52	0.50	0.59	0.75	0.85	0.85	0.92
264	0.67	0.67	0.50	0.56	0.50	0.59	0.42	0.51	0.25	0.38	0.92	0.96	1.00	1.00
267.12	0.33	0.53	0.33	0.44	0.67	0.77	0.75	0.84	0.75	0.80	1.00	1.00	1.00	1.00
267.5	0.67	0.78	0.75	0.85	0.83	0.90	0.67	0.76	0.50	0.62	1.00	1.00	1.00	1.00
267.8	0.47	0.54	0.56	0.62	0.66	0.69	0.56	0.63	0.60	0.66	0.83	0.87	0.83	0.88
268	0.45	0.54	0.25	0.41	0.35	0.44	0.35	0.44	0.40	0.49	0.50	0.65	0.75	0.86
279	0.83	0.86	0.92	0.92	0.75	0.81	0.83	0.88	0.83	0.86	1.00	1.00	0.92	0.96
380	0.24	0.35	0.24	0.36	0.39	0.47	0.47	0.53	0.35	0.48	0.78	0.80	0.71	0.73
462.37	0.40	0.49	0.40	0.52	0.65	0.69	0.67	0.70	0.65	0.69	0.78	0.80	0.81	0.87
465	0.50	0.63	0.75	0.76	0.50	0.63	0.38	0.54	0.75	0.75	1.00	1.00	1.00	1.00
467.1	0.29	0.41	0.57	0.75	0.67	0.76	0.33	0.64	0.58	0.70	1.00	1.00	1.00	1.00
495	0.32	0.40	0.32	0.46	0.60	0.66	0.56	0.61	0.60	0.65	0.77	0.87	0.87	0.92
530	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.03	1.00	1.00	1.00	1.00
591	0.13	0.31	0.25	0.36	0.61	0.71	0.52	0.58	0.51	0.60	0.87	0.93	0.87	0.92
601	0.58	0.62	0.58	0.64	0.86	0.89	0.79	0.81	0.29	0.49	0.86	0.93	0.86	0.93
650	0.64	0.70	0.72	0.77	0.78	0.80	0.69	0.74	0.75	0.76	0.97	0.99	0.97	0.99
672.73	0.25	0.30	0.25	0.32	0.33	0.34	0.33	0.34	0.33	0.38	0.67	0.71	1.00	1.00
672.78	0.27	0.34	0.34	0.43	0.42	0.46	0.50	0.55	0.42	0.48	0.83	0.92	1.00	1.00
676	0.14	0.29	0.14	0.27	0.50	0.62	0.57	0.66	0.36	0.55	0.93	0.94	1.00	1.00
683	0.11	0.26	0.18	0.30	0.48	0.52	0.52	0.57	0.48	0.54	0.81	0.88	0.90	0.94
684	0.35	0.43	0.60	0.72	0.25	0.51	0.25	0.34	0.25	0.27	1.00	1.00	1.00	1.00
686	0.21	0.28	0.28	0.36	0.57	0.65	0.65	0.68	0.43	0.55	0.68	0.79	0.94	0.96
687	0.20	0.30	0.30	0.49	0.62	0.64	0.38	0.51	0.50	0.53	0.88	0.94	0.75	0.83
715.1	0.12	0.25	0.12	0.22	0.33	0.50	0.33	0.45	0.50	0.56	1.00	1.00	1.00	1.00
718.1	0.17	0.26	0.08	0.24	0.67	0.67	0.67	0.67	0.33	0.48	1.00	1.00	1.00	1.00
718.2	0.20	0.30	0.17	0.31	0.52	0.59	0.52	0.57	0.59	0.64	0.76	0.85	0.87	0.92
784	0.20	0.29	0.30	0.49	0.50	0.52	0.38	0.46	0.50	0.52	1.00	1.00	1.00	1.00
839	0.17	0.25	0.07	0.20	0.33	0.37	0.33	0.36	0.33	0.35	1.00	1.00	1.00	1.00
Average	0.36	0.45	0.40	0.50	0.53	0.59	0.45	0.54	0.46	0.53	0.77	0.83	0.86	0.90

Table 15: Results on the 'Criminal Code Sections (Canada)' LEGALLAMA task. We kept only the sections with more than one example.

ECHR Article	Masked Terms
Art. 2	‘accessibility’ , ‘effective investigation’ , ‘expulsion’ , ‘extradition’ , ‘foreseeability’ , ‘positive obligations’ , ‘prescribed by law’ , ‘right to life’ , ‘safeguards against abuse’ , ‘use of force’
Art. 3	‘effective investigation’ , ‘expulsion’ , ‘extradition’ , ‘inhuman punishment’ , ‘inhuman treatment’ , ‘positive obligations’ , ‘prohibition of torture’ , ‘torture’
Art. 5	‘competent court’ , ‘deprivation of liberty’ , ‘drug addicts’ , ‘educational supervision’ , ‘expulsion’ , ‘extradition’ , ‘guarantees to appear for trial’ , ‘lawful arrest or detention’ , ‘lawful order of a court’ , ‘length of pre-trial detention’ , ‘minors’ , ‘order release’ , ‘persons of unsound mind’ , ‘procedure prescribed by law’ , ‘reasonable suspicion’ , ‘release pending trial’ , ‘review by a court’ , ‘right to liberty and security’ , ‘security of person’ , ‘speediness of review’ , ‘take proceedings’ , ‘trial within a reasonable time’
Art. 6	‘charged with a criminal offence’ , ‘disciplinary proceedings’ , ‘enforcement proceedings’ , ‘equality of arms’ , ‘examination of witnesses’ , ‘exclusion of public’ , ‘expulsion’ , ‘extradition’ , ‘fair hearing’ , ‘free legal assistance’ , ‘impartial tribunal’ , ‘independent tribunal’ , ‘insufficient means’ , ‘legal aid’ , ‘national security’ , ‘necessary in a democratic society’ , ‘oral hearing’ , ‘presumption of innocence’ , ‘protection of public order’ , ‘proved guilty according to law’ , ‘public hearing’ , ‘public judgment’ , ‘reasonable time’ , ‘right to a fair trial’ , ‘rights of defence’ , ‘same conditions’ , ‘tribunal established by law’
Art. 7	‘criminal offence’ , ‘heavier penalty’ , ‘retroactivity’
Art. 8	‘accessibility’ , ‘economic well-being of the country’ , ‘expulsion’ , ‘extradition’ , ‘foreseeability’ , ‘interference’ , ‘national security’ , ‘necessary in a democratic society’ , ‘positive obligations’ , ‘prevention of crime’ , ‘prevention of disorder’ , ‘protection of health’ , ‘protection of morals’ , ‘protection of the rights and freedoms of others’ , ‘public authority’ , ‘public safety’ , ‘respect for correspondence’ , ‘respect for family life’ , ‘respect for home’ , ‘respect for private life’ , ‘right to respect for private and family life’ , ‘safeguards against abuse’
Art. 9	‘foreseeability’ , ‘freedom of conscience’ , ‘freedom of religion’ , ‘freedom of thought’ , ‘interference’ , ‘necessary in a democratic society’ , ‘observance’ , ‘positive obligations’ , ‘practice’ , ‘prescribed by law’ , ‘protection of health’ , ‘protection of public order’ , ‘protection of the rights and freedoms of others’ , ‘public safety’ , ‘safeguards against abuse’ , ‘teaching’ , ‘worship’
Art. 10	‘duties and responsibilities’ , ‘foreseeability’ , ‘freedom of expression’ , ‘freedom to hold opinions’ , ‘freedom to impart information’ , ‘freedom to receive information’ , ‘interference’ , ‘national security’ , ‘necessary in a democratic society’ , ‘positive obligations’ , ‘prescribed by law’ , ‘prevention of crime’ , ‘prevention of disorder’ , ‘protection of health’ , ‘protection of morals’ , ‘protection of the reputation of others’ , ‘protection of the rights of others’ , ‘public safety’ , ‘safeguards against abuse’ , ‘territorial integrity’
Art. 11	‘accessibility’ , ‘foreseeability’ , ‘form and join trade unions’ , ‘freedom of assembly and association’ , ‘freedom of association’ , ‘freedom of peaceful assembly’ , ‘interference’ , ‘national security’ , ‘necessary in a democratic society’ , ‘positive obligations’ , ‘prescribed by law’ , ‘prevention of crime’ , ‘prevention of disorder’ , ‘protection of health’ , ‘public safety’
Art. 13	‘effective remedy’ , ‘national authority’ , ‘right to an effective remedy’
Art. 14	‘discrimination’ , ‘language’ , ‘national minority’ , ‘national origin’ , ‘objective and reasonable justification’ , ‘prohibition of discrimination’ , ‘property’ , ‘race’ , ‘religion’ , ‘sex’ , ‘social origin’
Art. 35	‘continuing situation’ , ‘effective domestic remedy’ , ‘exhaustion of domestic remedies’ , ‘final domestic decision’ , ‘manifestly ill-founded’ , ‘no significant disadvantage’ , ‘relevant new information’
Art. P1-1	‘accessibility’ , ‘deprivation of property’ , ‘foreseeability’ , ‘general interest’ , ‘general principles of international law’ , ‘interference’ , ‘peaceful enjoyment of possessions’ , ‘positive obligations’ , ‘possessions’ , ‘prescribed by law’ , ‘protection of property’ , ‘secure the payment of taxes’

Table 16: Masked Terms used in the ‘Terminology (CoE)’ LEGALLAMA task.

Crime Area	Masked Terms
Children	'child abandonment' , 'child abuse'
Computer	'computer crime' , 'cyberbullying' , 'identity theft'
Court-related	'criminal contempt of court' , 'perjury' , 'probation violation'
Drug-related	'drug distribution' , 'drug manufacturing' , 'drug possession' , 'drug trafficking' , 'medical marijuana' , 'minor in possession' , 'public intoxication'
Life Taking	'homicide' , 'manslaughter' , 'murder'
Mens Rea	'accessory' , 'aiding and abetting' , 'attempt' , 'conspiracy' , 'hate crime'
Monetary	'bribery' , 'embezzlement' , 'extortion' , 'forgery' , 'insurance fraud' , 'money laundering' , 'pyramid schemes' , 'racketeering' , 'securities fraud' , 'shoplifting' , 'tax evasion' , 'telemarketing fraud' , 'theft' , 'white collar crime' , 'wire fraud'
Behavior	'disorderly conduct' , 'disturbing the peace' , 'harassment' , 'stalking'
Property	'arson' , 'vandalism'
Sex-related	'child pornography' , 'indecent exposure' , 'prostitution' , 'rape' , 'sexual assault' , 'solicitation' , 'statutory rape'
Violence	'aggravated assault' , 'battery' , 'burglary' , 'domestic violence' , 'kidnapping' , 'robbery'

Table 17: Masked Terms used in the 'Crime Charges (US)' LEGALLAMA task grouped by crime areas.

Legal Topic	Masked Terms
Business Law	'adhesion contract' , 'implied warranty' , 'limited liability' , 'parol evidence' , 'quantum meruit' , 'reliance damages' , 'self-dealing' , 'severability clause' , 'specific performance' , 'statute of frauds' , 'substantial performance' , 'tender offer' , 'third-party beneficiary' , 'unconscionability'
Criminal Law and Procedure	'accessory before the fact' , 'accomplice' , 'aggravated assault' , 'allocution' , 'arson' , 'defense of others' , 'inchoate' , 'merger doctrine' , 'mitigating circumstances' , 'money laundering' , 'stop and frisk'
Employment Law	'bargaining unit' , 'boycott' , 'casual labor' , 'industrial safety' , 'minimum wage' , 'workplace safety' , 'wrongful termination'
Family Law	'consent divorce' , 'emancipation of minors' , 'marital privilege' , 'marital property' , 'marital settlement agreement' , 'separate property' , 'separation agreement' , 'shared custody' , 'sole custody' , 'spousal privilege' , 'spousal support' , 'visitation' , 'wage attachment'
Immigration	'alienage' , 'asylum seeker' , 'asylum' , 'childhood arrivals' , 'citizenship' , 'deferred action' , 'deportation' , 'geneva conventions' , 'naturalization' , 'nonresident' , 'refugee' , 'resettlement' , 'visa'
Landlord-Tenant Law	'abandonment' , 'commercial reasonability' , 'constructive eviction' , 'eviction' , 'habitability' , 'privity' , 'quiet enjoyment' , 'reasonableness' , 'self-help eviction' , 'sole discretion' , 'tenancy at sufferance' , 'tenancy at will'
Money And Financial Problems	'bankruptcy discharge' , 'bond' , 'consumer credit' , 'kiting' , 'malfeasance' , 'mortgage' , 'nonrecourse' , 'ponzi scheme' , 'securities fraud' , 'self-dealing' , 'senior lien' , 'stock dividend' , 'straw man' , 'swindle' , 'tontine' , 'variable annuity'

Table 18: Masked Terms used in the 'Terminology (US)' LEGALLAMA task grouped by legal topics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 2-3

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Section 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.