The Large Language Model GreekLegalRoBERTa

Vasileios Saketos*
saketos@kth.se
KTH Royal Institute of Technology
Stockholm , Sweden

Despina-Athanasia Pantazi*
dpantazi@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

Manolis Koubarakis*
koubarak@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

ABSTRACT

We develop four versions of GreekLegalRoBERTa, which are four large language models trained on Greek legal and nonlegal text. We show that our models surpass the performance of GreekLegalBERT, Greek- LegalBERT-v2, and GreekBERT in two tasks involving Greek legal documents: named entity recognition and multi-class legal topic classification. We view our work as a contribution to the study of domain-specific NLP tasks in low-resource languages, like Greek, using modern NLP techniques and methodologies.

KEYWORDS

Natural Language Processing, Pre-trained Language Models, Greek NLP Resources, Greek Legislation, Classification, Named Entity Recognition

ACM Reference Format:
Vasileios Saketos, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2024. The Large Language Model GreekLegalRoBERTa. In *13th Conference on Artificial Intelligence (SETN 2024), September 11–13, 2024, Piraeus, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3688671.3688770

1 INTRODUCTION

The success of the Transformer architecture [32] has led to the creation of many large language models (LLMs) and related benchmarks for the legal domain [7–9, 24]. Of particular interest to us is the development of legal LLMs and benchmarks for low-resource languages.

For the Greek language in particular, [1] studied the problem of named entity recognition (NER) in Greek legal documents using various kinds of RNN networks. The paper also developed the dataset GreekLegalNER[1] which has been constructed using Greek legislation available in the platform Nomothesia [10]. Nomothesia is a platform that makes Greek legislation available on the Web as linked data using appropriate legal ontologies.

Later on, [27] proposed GreekLegalBERT, a version of BERT trained on Greek legislation available in the platform Nomothesia [10]. In [27], GreekLegalBERT was applied to the task of multi-class legal topic classification using the dataset GreekLegalCode consisting of 47k legal documents from Greek legislation.

[1]This name is not used in [1]; it is the name [24] gave to this dataset (part of their LEXTREME benchmark), so we use the same name for consistency.

SETN 2024, September 11–13, 2024, Piraeus, Greece

ACM ISBN 979-8-4007-0982-1/24/09.
https://doi.org/10.1145/3688671.3688770

In this work, we explore the use of RoBERTa [23] in the tasks of NER and multi-class legal topic classification for Greek legislation. RoBERTa is a version of BERT [12] that has been obtained using a simpler training method which, nevertheless, leads to a significant gain in performance. The contributions of our paper are the following:

- We develop four versions of GreekLegalRoBERTa. These four different LLMs were trained on the dataset available in the Nomothesia platform [10], the Greek Parliament Proceedings [13], the Greek version of European Parliament Proceedings Parallel Corpus [19], the Greek part of European Union legislation [6], Raptarchis [27] and OSCAR [26]. We view our work as a contribution to the study of domain-specific NLP tasks in low-resource languages.
- We apply the trained models on the tasks of GreekLegalNER in Greek legal documents originally studied in [1], and multi-class legal topic for the GreekLegalCode dataset presented in [27]. Our models managed to surpass the performance of all the previous models in micro and weighted average in GreekLegalNER and all the tasks of GreekLegalCode. More specifically, we managed to improve performance over the state-of-the-art results obtained by Greek- LegalBERT-v2 by 1.2 points in micro average and 1.4 points in weighted average within GreekLegalNER. Moreover, we managed to improve by 0.12 points in volume, 0.89 points in chapters and 0.67 points in subjects of GreekLegalCode.
- We make our models, datasets, and code publicly available [2] so that they can be used by the research community.

The organization of the rest of the paper is as follows. Section 2 discusses related work. Section 3 analyzes the current state-of-the-art models and the different configurations we used to pretrain our models, while in Section 4, we apply the developed models to the tasks of NER and multi-class legal topic classification. Section 5 concludes the paper by discussing limitations and future work.

2 RELATED WORK

If one wants to solve an NLP task for a low resource language like Greek, one solution is to use a multilingual LLM. For Greek, both M-BERT and XLM-RoBERTa [11] will do for the task, since their training corpora includes Greek documents. The first monolingual language model to be proposed for the Greek language is GreekBERT [20]. It has been applied to the tasks of part-of-speech (POS) tagging, NER and natural language inference (NLI), and it was shown to outperform M-BERT and XLM-RoBERTa on these tasks.

[2]Our code is available at : https://github.com/AI-team-UoA/GreekLegalRoBERTa
Our models and datasets are available at: https://huggingface.co/AI-team-UoA

Nowadays, there has been a notable emphasis on the development and advancement of generative multilingual models. Some outstanding examples of these models are OPT[34], BLOOM [28] and GPT3 [5], while the most recent and prominent are PaLM[4], Chinchilla[16] and Llama[31]. These models demonstrate significant performance in both natural and programming languages. While OPT, BLOOM, and Llama are open-source models trained on publicly available data, GPT-3, Chinchilla, and PaLM are closed source and trained on private datasets. Very recently, the generative LLM Meltemi has been developed for the Greek language [3]. Nevertheless, considering that each model comprises hundreds of billions of parameters, the resources required to utilize these models are substantial.

Moreover, in the generative question answering task, the responses of these models suffer from extensive hallucination. The term hallucination refers to the phenomenon where a model generates information that is incorrect, misleading, or entirely fabricated. Retrieval Augmented Generation (RAG) [22] emerged to address this issue. RAG architectures consist of a retriever p_η and a generator p_θ. The generator is an encoder decoder architecture such as BART [21], Llama [31], and BLOOM [28]. The retriever is an encoder architecture such as DPR [18], E5 [33] and Nomic [25], and it is responsible for retrieving the top-k passages $(y_1, y_2, ..., y_k)$ from a database given a query x_i. Then, the query along with the k passages are passed to the generator. Finally, the generator produces the response to the original query. Despite differences in the fine-tuning methods and architectures of the retrievers, they leverage the same objective. Their objective is to ensure that the cosine_similarity$(p_\eta(x_i), p_\eta(y_j))$ accurately reflects the true relationship between the query x_i and the passage y_j. To achieve this, they start from a pre-trained encoder model like RoBERTa and they fine-tune it according to this objective. Empirical results [30] have shown that RAG reduces hallucination to a great extent.

Regarding the legal domain, the first language model proposed for Greek legislation is GreekLegalBERT [12], as we already mentioned in the introduction. It was applied to the task of multi-class legal topic classification for the dataset GreekLegalCode and it was shown to consistently outperform traditional machine learning algorithms (SVM and XG-BOOST), BiGRU-based methods and the multilingual models M-BERT and XLM-RoBERTa on this task.

With respect to the available legal benchmarks, the paper [15] presents LEGALBENCH, a collaboratively constructed legal reasoning benchmark which consists of 162 tasks covering six different types of legal reasoning. LEGALBENCH is important because legal professionals had a leading role in its construction. LEGALBENCH contains 112 legal binary classification tasks and 8 multi-class classification tasks. Recently, [24] proposed LEXTREME, a benchmark consisting of 11 legal datasets covering 24 languages and 18 tasks. They applied five popular encoder-based LLMs and found that the model size correlates with the performance on the benchmark in most cases. From the five models evaluated, XLM-RoBERTa is the most effective one. The datasets GreekLegalNER and GreekLegalCode studied in this paper are part of LEXTREME. What distinguishes our work from larger efforts such as LEXTREME is that by concentrating on a single language, we manage to develop language models that are more effective than the multilingual ones used in [24]. This paper reports on how we have achieved this using a version of RoBERTa trained on Greek legal documents.

3 THE GREEKLEGALROBERTA MODELS

In this work, we focus on the use of RoBERTa and we develop four variations of our model GreekLegalRoBERTa. In the rest of the section, we provide the detailed description of the models in comparison to the existing state-of-the-art Greek models GreekBERT [20], GreekLegalBERT [3] and GreekLegalBERT-v2 [2]. In Table 1, we present the statistics of the pretraining corpora used in the various models.

GreekBERT [20]: This model is a monolingual version of BERT, trained solely on modern Greek, achieving state-of-the-art results in most of the Greek NLP tasks. BERT models are pretrained for 1M steps and batch size of 256. To speed up the pretraining process they used a sequence length of 128 for 90% of the training steps. Then, for the remaining 10% of steps, they used a sequence length of 512 to learn the positional embeddings. During pretraining the objective of BERT is to maximize the performance in Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [17]. GreekBERT was pretrained on 29GB of text from a corpus consisting of the Greek part of Wikipedia, the Greek part of the European Parliament Proceedings Parallel Corpus(Europarl) [19] and OSCAR [26].

GreekLegalBERT [3]: This model is a monolingual legal version of BERT trained on dataset accessible through the Nomothesia [10] platform. The dataset consists of laws, announcements, and resolutions in the Greek language. The total size of the dataset is 5GB and it spans a chronological range from 1990 to 2017. Despite the smaller dataset, this model managed to exceed the performance of GreekBERT in GreekLegalCode and match its performance GreekLegalNER.

GreekLegalBERT-v2 [2]: This model is a more recent version of GreekLegalBERT pretrained on the Nomothesia dataset, the Greek Parliament Proceedings (Greekparl) [13], Eurparl [19], the Greek part of European Union legislation (Eurlex) [6] and Raptarchis [27]. The total size of the dataset sums up to 8GB.

GreekLegalRoBERTa-v1: This is the first model we train for the purposes of our work. This model is pretrained on the dataset accessible through the Nomothesia [10] platform.

The distinctive characteristics of our dataset render the preprocessing of the data essential. Due to the rather large chronological range of our data, a wide range of character encodings have been used. Windows-1253 and ISO 8859-7 were among the character encodings utilized. We therefore need to convert all identical characters into a unique representation. Moreover, we normalize our data using the normalization form compatibility decomposition (NFKD), because K normalizations are more effective in eliminating formatting distinctions. Additionally, we remove accents due to the possibility of words in Greek having the same letters but differing in their accents.

To encode our text we utilize Byte-Pair Encoding (BPE)[29], a hybrid between character and word level representations that allows handling the vocabularies common in natural language corpora.

[3]https://medium.com/institute-for-language-and-speech-processing/meltemi-a-large-language-model-for-greek-9f5ef1d4a10f

Instead of full words, BPE relies on subwords units, which are extracted by performing statistical analysis of the training corpus. To train our encoder, we utilize a vocabulary of size 50 264. BPE training was done in the same dataset used to pretrain our model.

In contrast to BERT, RoBERTa utilizes a sequence length of 512 throughout the entire pretraining process. This modification significantly increases the computational cost, as the attention mechanism's computational complexity grows quadratically with the sequence length. To avoid using the same mask for each training instance in every epoch in BERT, training data was duplicated 10 times so that each sequence is masked in 10 different ways over the 40 epochs of training. On the contrary, RoBERTa uses dynamic masking. In dynamic masking, the masking pattern is generated every time we feed a sequence to the model. In the original RoBERTa paper, it was concluded that dynamic masking performs slightly better than static masking. By utilizing dynamic masking, we avoid duplication and consequently we reduce significantly the memory requirements during training. Additionally, the authors removed the NSP objective as it was shown that using the NSP objective hurts the performance on downstream tasks. The mixed floating point precision was used to train the original RoBERTa model. This technique increases stability during training while speeding up the training process.

We pretrain for 100k steps and a batch size of 1024. This experiment solidifies the statement that pretraining RoBERTa leads to more prominent results even when we pretrain less compared to BERT. The model was pretrained using a single GPU. The total training duration of this model amounted to 30 days of uninterrupted training.

GreekLegalRoBERTa-v2: This model is identical to v1 except for the fact that we trained our model with a batch size of 4096 in a 4 GPU V100 cluster. In this model, our endeavor is to leverage the principle that within modern Natural Language Processing, more pretraining of a model correlates with enhanced performance. Moreover, training with large batches improves the perplexity for the masked language modeling objective, as well as the end-task accuracy. The total training duration of this model amounted to 40 days of uninterrupted training.

GreekLegalRoBERTa-v3: In our experiments discussed in section 4, it is evident that GreekLegalBERT-v2 exhibits improved performance compared to GreekLegalBERT. Our objective is to enhance the performance of our new model by leveraging the supplementary dataset employed in GreekLegalBERT-v2. We adopt the identical experimental setup as employed in GreekLegalRoBERTa-v2.

GreekLegalRoBERTa-v4: Last but not least, we utilize all the datasets we discussed in the previous models. In this model, we aim to ascertain whether incorporating both legal and non-legal contexts results in improved performance on legal tasks. We adopt the identical experimental setup as employed in GreekLegalRoBERTa-v2.

Having presented the models, we can now proceed to apply them and evaluate their performance.

4 EXPERIMENTS

We compare the performance of our newly introduced four GreekLegalRoBERTa models to the existing state-of-the-art Greek models GreekBERT, GreekLegalBERT and GreekLegalBERT-v2, against baselines on datasets for two core downstream tasks: NER and multi-class legal topic classification. In Tables 2 and 3, we present our results. We observe that our models provide improvement over the originally reported performance of all the compared models.

Table 1: Statistics of the pretraining corpora

Corpus	**Size (GB)**	**Context**
Wikipedia	1.73	nonlegal
Nomothesia dataset	5	legal
Europarl	0.38	legal
Eurolex	0.41	legal
Greekparl	2.7	legal
Raptrarchis	0.22	legal
OSCAR	27	nonlegal
Total	37.03	-

4.1 GreekLegalNER

For the first downstream task, NER, we test the models against the benchmark GreekLegalNER, introduced by [1].

Dataset: The dataset contains 254 daily issues of the Greek Government Gazette over the period 2000-2017. Every issue contains multiple legal acts. This dataset focuses on 7 entity types: legislation references, geopolitical entities, national locations, unknown locations, public locations, organizations, and facilities. The dataset is available in the Inside Outside Beginning (IOB) format, which is a common tagging format for tagging tokens for NER. It consists of 35 411 instances and it is divided into 3 main parts: *train* 67.5%, *validation* 17.5%, and *test* 15%.

Evaluation: We experiment with epochs from 1 to 20, batch size 8 and 16, learning rate 2e-5, 3e-5 and 5e-5. Table 5 presents the hyperparameter combinations with which the models achieved their best results. Then, we perform 5 runs of finetuning and model evaluation per model using 5 different *seeds*. For each model's performance on the *test* set, we present the mean F1 score of every entity type, micro, macro and weighted F1, including the standard deviation of the 5 experiments in Table 2. In order to perform a comparative evaluation, we highlight the best F1 performance evaluation for each entity.

Our model GreekLegalRoBERTa-v2 surpasses the performance of the state of the art GreekLegalBERT-v2 on micro F1 by 1.2 points and on weighted F1 by 1.4 points. Nevertheless, GreekBERT archives greater performance by 0.8 points than GreekLegalRoBERTa-v2. It is also evident that the RoBERTa models demonstrate superior performance compared to the BERT models trained on the same dataset.

Based on Table 2, GreekBERT achieves the highest score in 3 out of the 8 entity types, GreekLegalBERT in 0 out of 8, GreekLegalBERT-v2 in 1 out of 8, GreekLegalRoBERTa-v1 in 2 out of 8, GreekLegalRoBERTa-v2 in 5 out of 8, GreekLegalRoberta-v3 in 3 out of 8, and

Table 2: Results on the task of NER using F1 scores

	F	GPE	LR	LN	LU	ORG	P	PD	micro	macro	weighted
GreekBERT	31 (3%)	**75 (1%)**	82 (0%)	**88 (11%)**	**73 (1%)**	73 (1%)	85 (1%)	68 (1%)	76.2	**71.8**	76.2
GreekLegalBERT	29 (3%)	74 (1%)	82 (0%)	43 (16%)	71 (2%)	76 (1%)	86 (1%)	68 (2%)	76	66	76
GreekLegalBERT-v2	31 (4%)	**75 (1%)**	82 (0%)	62 (1%)	72 (1%)	76 (1%)	87 (1%)	67 (4%)	76.4	69	76.2
GreekLegalRoBERTa-v1	30 (2%)	**75 (0%)**	82 (1%)	55 (20%)	71 (1%)	**77 (1%)**	87 (1%)	68 (0%)	76.8	68	76.8
GreekLegalRoBERTa-v2	30 (2%)	**75 (1%)**	**83 (1%)**	73 (19%)	72 (1%)	**77 (1%)**	**88 (1%)**	**70 (0%)**	**77.6**	71	**77.6**
GreekLegalRoBERTa-v3	33 (1%)	74 (1%)	**83** (0%)	69 (10%)	71 (1%)	76 (1%)	**88 (1%)**	**70 (1%)**	76.8	70.5	76.8
GreekLegalRoBERTa-v4	**35 (2%)**	74 (1%)	82 (1%)	67 (2%)	**73 (1%)**	77 (1%)	87 (1%)	68 (1%)	77	71	77

To accommodate the results within the confines of the page, the utilization of the following acronyms is employed: Facility: F, Geopolitical Entity: GPE, Legislation Reference: LR, National Location: LN, Unknown Location: LU, Organization: ORG, Person: P, Public Document: PD.

Table 3: Results on the task of multi-class legal topic classification

	Volume			**Chapter**			**Subject**		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
GreekBERT	89.84	89.84	89.84	84.87	84.87	84.87	80.59	80.59	80.59
GreekLegalBERT	90.51	90.51	90.51	85.45	85.45	85.45	81.43	81.43	81.43
GreekLegalBERT-v2	91.02	91.02	91.02	85.72	85.72	85.72	82.71	82.71	82.71
GreekLegalRoBERTa-v1	91.10	91.10	91.10	85.77	85.77	85.77	82.29	82.29	82.29
GreekLegalRoBERTa-v2	**91.14**	**91.14**	**91.14**	86.29	86.29	86.29	82.72	82.72	82.72
GreekLegalRoBERTa-v3	89.04	89.04	89.04	**86.61**	**86.61**	**86.61**	**83.38**	**83.38**	**83.38**
GreekLegalRoBERTa-v4	91.02	91.02	91.02	86.31	86.31	86.31	82.54	82.54	82.54

GreekLegalRoberta-v4 in 2 out of 8. It is also clear that the models trained on non-legal contexts shown in Table 4 perform better in Facility entity type and National and Unknown Location types. This can occur because the datasets used during pretraining contain more instances of Facilities and Locations. On the other hand, models trained only on legal context tend to perform better in Person, Organization, and Public document entity types.

Furthermore, our anticipation was that the utilization of a larger dataset for training GreekLegalRoberta-v4 would result in a performance that significantly outperforms all the preceding models. However, contrary to our expectations, this was not the case. This could be attributed to the substantial difference between the vocabulary employed in the Greek legal written context and that utilized in everyday life. As a consequence, our tokenizer inadequately represents the expanded vocabulary, leading to suboptimal performance.

4.2 GreekLegalCode

Lastly, for the second downstream task, multi-class legal topic classification, we conduct experiments using the GreekLegalCode dataset[27]. The GreekLegalCode paper introduces a new dataset of legal context and proves that using this dataset, GreekLegalBERT outperforms all the previous Greek and multilingual models. In Table 3, we show that our models outperform all the previous ones including the GreekLegalBERT versions.

Dataset: The dataset is a thorough classification of the Greek legislation. It includes Laws, Royal and Presidential Decrees, Regulations, and Decisions, retrieved from the Official Government Gazette. The dataset is structured into thematic topics, making the data ideal for multi-label classification. It consists of 47 legislative *volumes* and each *volume* corresponds to a main thematic topic. Each *volume* is divided into thematic subcategories which are called *chapters* and subsequently, each *chapter* breaks down into *subjects*. The total number of *chapters* is 389 while the total number of *subjects* is 2285.

Evaluation: As proposed in the original paper[27], we perform grid-search over the core hyper-parameters. More specifically, we experiment with epochs from 1 to 20 and learning rate 1e-5, 2e-5, 3e-5, and 5e-5. We present the best hyperparameter configuration in Table 5. The experimental setup is the same as in GreekLegalNER, analyzed in the previous subsection. We present the mean of *micro* F1, precision, and recall of the performance evaluation in Table 3. We start from the established benchmark obtained by GreekLegalBERT-v2 which is 91.02 for Volume, 85.72 for Chapter, and 82.71 for Subject. Our model GreekLegalRoBERTa-v2 managed to achieve greater performance by 0.12 points on Volume. Moreover GreekLegalRoBERTa-v3 surpasses the performance of GreekLegalBERT-v2 by 0.89 on chapter and 0.67 on subject.

5 CONCLUSION AND FUTURE WORK

In this work, we introduce four new language models pretrained in legal and nonlegal text. We have attained state-of-the-art results across all the tasks of the GreekLegalCode dataset. Moreover, our models demonstrate superior performance compared to all the other models across all metrics, with the exception of the macro average in GreekLegalNER. Among the two conducted experiments, GreekLegalNER emerges as the more challenging task. One reason for this can be that the models have difficulty classifying correctly the two different types of location entities.

Finally, for future work, it would be interesting to see how well the models perform if we combine the two different location categories. Moreover, we aim to develop more Natural Language Processing models in modern Greek and legal contexts, as we believe it is crucial, especially in low-resource languages like Greek. By providing NLP models in specific and non-specific domains not only do we create opportunities and tools for future research, but we also help the industry to provide more reliable products to everyone.

Furthermore, nowadays MergeKit [14] provides the opportunity to merge different NLP models of the same architectures in any hardware. By combining two NLP models, the new model can utilize each other's strengths without additional pretraining, offering a promising avenue for enhancing a model's performance. Developing additional NLP models for specific and non specific context is essential to fully capitalize on this opportunity.

Last but not least, our model can be used as a foundation for developing high-quality retrievers like DPR [18], and E5 [33], specifically for Greek legislation. By training encoder decoder architecture like BART [21], or Llama [31] in Greek legal text, we can establish a highly effective architecture for Retrieval Augmented Generation (RAG) tailored to the Greek legislation.

6 ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the Cyprus Research Institute for providing us with access to Cyclone. Cyclone is a cluster which consists of 16 40-core compute nodes and 4 Nvidia V100 GPUs each. This work would not have been possible without this contribution.

REFERENCES

[1] Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *Legal Knowledge and Information Systems - JURIX 2018: The Thirty-first Annual Conference, Groningen, The Netherlands, 12-14 December 2018 (Frontiers in Artificial Intelligence and Applications, Vol. 313)*, Monica Palmirani (Ed.). IOS Press, 1–10. https://doi.org/10.3233/978-1-61499-935-5-1

[2] Alexandra Apostolopoulou and Spyridon Mpriakos. 2022. NLP Tasks with GreekLegalBERT v2. https://pergamos.lib.uoa.gr/uoa/dl/frontend/el/browse/2971631. Undergraduate dissertation. Dept. Informatics and Telecommunication, National and Kapodistrian University of Athens.

[3] Konstaninos Athinaios. 2020. Named Entity Recognition using a Novel Linguistic Model for Greek Legal Corpora based on BERT model. https://pergamos.lib.uoa.gr/uoa/dl/object/2927727. Undergraduate dissertation. Dept. Informatics and Telecommunication, National and Kapodistrian University of Athens.

[4] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. PALM: Pre-training an Autoencoding&Autoregressive Language Model for Context-conditioned Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8681–8691. https://doi.org/10.18653/v1/2020.emnlp-main.700

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[6] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. 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*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6974–6996. https://doi.org/10.18653/v1/2021.emnlp-main.559

[7] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: "Preparing the Muppets for Court'". In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[8] Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. *CoRR* abs/2305.07507 (2023). https://doi.org/10.48550/arXiv.2305.07507 arXiv:2305.07507

[9] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

[10] Ilias Chalkidis, Charalampos Nikolaou, Panagiotis Soursos, and Manolis Koubarakis. 2017. Modeling and Querying Greek Legislation Using Semantic Web Technologies. In *The Semantic Web*, Eva Blomqvist, Diana Maynard, Aldo Gangemi, Rinke Hoekstra, Pascal Hitzler, and Olaf Hartig (Eds.). Springer International Publishing, Cham, 591–606.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[12] 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Konstantina Dritsa, Aikaterini Thoma, Ioannis Pavlopoulos, and Panos Louridas. 2022. A Greek Parliament Proceedings Dataset for Computational Linguistics and Political Analysis. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 28874–28888. https://proceedings.neurips.cc/paper_files/paper/2022/file/b96ce67b2f2d45e4ab315e13a6b5b9c5-Paper-Datasets_and_Benchmarks.pdf

[14] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A Toolkit for Merging Large Language Models. arXiv:2403.13257 [cs.CL]

[15] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. arXiv:2308.11462 [cs.CL]

[16] 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. *CoRR* abs/2203.15556 (2022). https://doi.org/10.48550/ARXIV.2203.15556 arXiv:2203.15556

[17] Yacine Jernite, Samuel R. Bowman, and David A. Sontag. 2017. Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning. *CoRR* abs/1705.00557 (2017). arXiv:1705.00557 http://arxiv.org/abs/1705.00557

[18] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. *ArXiv* abs/2004.04906 (2020). https://api.semanticscholar.org/CorpusID:215737187

[19] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. https://aclanthology.org/2005.mtsummit-papers.11. In *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, 79–86.

[20] John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. https://dl.acm.org/doi/10.1145/3411408.3411440. In *SETN 2020: 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, September 2-4, 2020*, Constantine D. Spyropoulos, Iraklis Varlamis, Ion Androutsopoulos, and Prodromos Malakasiotis (Eds.). ACM, 110–117. https://dl.acm.org/doi/10.1145/3411408.3411440

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[23] 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 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[24] 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. *CoRR* abs/2301.13126 (2023). https://doi.org/10.48550/arXiv.2301.13126 arXiv:2301.13126

[25] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic Embed: Training a Reproducible Long Context Text Embedder. *ArXiv* abs/2402.01613 (2024). https://api.semanticscholar.org/CorpusID:267406738

[26] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. https://www.aclweb.org/anthology/2020.acl-main.156. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1703–1714.

[27] Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular Legal Topic Classification on Greek Legislation. https://aclanthology.org/2021.nllp-1.6". In *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 63–75. https://doi.org/10.18653/v1/2021.nllp-1.6

[28] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR* abs/2211.05100 (2022). https://doi.org/10.48550/ARXIV.2211.05100 arXiv:2211.05100

[29] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1715–1725. https://doi.org/10.18653/v1/P16-1162

[30] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:233240939

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

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[33] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv:2212.03533 [cs.CL] https://arxiv.org/abs/2212.03533

[34] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *CoRR* abs/2205.01068 (2022). https://doi.org/10.48550/ARXIV.2205.01068 arXiv:2205.01068

A APPENDIX

A.1 Pretraining hyperparameters

In Table 4, we provide an overview of the hyperparameters utilized during the pretraining phase for all the models.

A.2 Finetuning hyperparameters

In Table 5, we present the optimal hyperparameters for both the GreekLegalCode and GreekLegalNER tasks.

Table 4: Detailed comparison of the models' pretraining

	Training steps	Batch size	Vocabulary size	Size	Context
GreekBERT	1M	256	35k	29GB	nonlegal
GreekLegalBERT	1M	256	35k	5GB	legal
GreekLegalBERT-v2	1M	256	35k	8GB	legal
GreekLegalRoBERTa-v1	100K	1024	50k	5GB	legal
GreekLegalRoBERTa-v2	100K	4096	50k	5GB	legal
GreekLegalRoBERTa-v3	1M	256	50k	8GB	legal
GreekLegalRoBERTa-v4	1M	256	50k	37GB	legal and nonlegal

Table 5: Best hyperparameters on finetuning the models for the experiments

	volume		chapter		subject		NER		
Model	**learning rate**	**epochs**	**learning rate**	**epochs**	**learning rate**	**epochs**	**learning rate**	**epochs**	**batch size**
GreekBERT	5e-5	3	3e-5	5	5e-5	15	5e-5	3	8
GreekLegalBERT	3e-5	3	5e-5	5	3e-5	15	3e-5	3	8
GreekLegalBERT-v2	3e-5	3	3e-5	8	3e-5	13	3e-5	3	8
GreekLegalRoBERTa-v1	5e-5	6	3e-5	5	5e-5	15	5e-5	6	8
GreekLegalRoBERTa-v2	5e-5	6	5e-5	6	5e-5	14	5e-5	6	8
GreekLegalRoBERTa-v3	5e-5	5	5e-5	13	5e-5	24	5e-5	3	8
GreekLegalRoBERTa-v4	3e-5	7	5e-5	21	5e-5	18	3e-5	7	8