Large Language Models are legal but they are not: Making the case for a powerful LegalLLM

Thanmay Jayakumar, Fauzan Farooqui*, Luqman Farooqui*
Visvesvaraya National Institute of Technology, Nagpur, India
{thanmayjayakumar, fauzanfarooqui7, luqmanfarooqui99}@gmail.com

Abstract

Realizing the recent advances in Natural Language Processing (NLP) to the legal sector poses challenging problems such as extremely long sequence lengths, specialized vocabulary that is usually only understood by legal professionals, and high amounts of data imbalance. The recent surge of Large Language Models (LLM) has begun to provide new opportunities to apply NLP in the legal domain due to their ability to handle lengthy, complex sequences. Moreover, the emergence of domain-specific LLMs has displayed extremely promising results on various tasks. In this study, we aim to quantify how general LLMs perform in comparison to legal-domain models (be it an LLM or otherwise). Specifically, we compare the zero-shot performance of three general-purpose LLMs (ChatGPT-3.5, LLaMA-2-70b, and Falcon-180b) on the LEDGAR subset of the LexGLUE benchmark for contract provision classification. Although the LLMs were not explicitly trained on legal data, we observe that they are still able to classify the theme correctly in most cases. However, we find that their mic-F1/mac-F1 performance is upto 19.2/26.8% lesser than smaller models fine-tuned on the legal domain, thus underscoring the need for more powerful legal-domain LLMs.

1 Introduction

Legal professionals typically deal with large amounts of textual information on a daily basis to make well-informed decisions in their practice. This can become very tedious and demanding due to the overwhelming amount of data they must manage and the meticulous attention to detail necessary to maintain the required precision in their work. Thanks to the rise of LLMs, many tasks such as sentiment analysis, named entity recognition, information retrieval, etc. can now be handled by neural models. Though this holds true for the legal domain as well (Sun, 2023), they aren't used to make direct decisions. Nevertheless, these automated systems that produce legal predictions and generations, are predominantly useful as advisory tools for legal practitioners that can augment their decision-making process.

Transformers (Vaswani et al., 2017) have become the *de facto* method for many text classification and multiple choice question answering tasks. BERT (Devlin et al., 2019), a transformer-encoder, and its derived models like RoBERTa (Liu et al., 2019) are commonly employed in legal NLP tasks. Pre-training such models on legal corpora can help a model adapt to a specific domain by fine-tuning it with domain-specific data. LegalBERT (Chalkidis et al., 2020) is one such BERT model that was trained on legal-oriented data. CaseLaw-BERT (Zheng et al., 2021), PoL-BERT (Henderson et al., 2022), and LexLM (Chalkidis et al., 2023) are a few more BERT-based variants pre-trained for the legal domain. Although they show remarkable performance on various legal tasks in comparison with general-purpose BERT models, one limit of these models is that BERT's input size can only incorporate a maximum of 512 tokens. For short sequences this may seem enough, but in the case of long documents which is commonly found in the legal domain, where input texts can go over 5000 tokens (and requiring *even* more in few-shot settings), it can be a severe drawback as a lot of important information will get truncated.

Due to this limit, BERT-based models aren't employed as-is in long-document tasks. Typically, methods like hierarchical attention are utilized where the long document is split into segments of max length (512 in the case of BERT models) and these segments are independently encoded. These segment embeddings are then aggregated with stacked transformers to get the overall encoding of the entire document. Similarly, recurrent

*These authors contributed equally

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 223–229

transformers (Dai et al., 2019; Yang et al., 2019; Ding et al., 2021) were proposed to process long documents by encoding its representation from individual segments in a recurrent fashion. Sparse attention is another method that has been proposed to tackle long sequence inputs (Ainslie et al., 2020; Zaheer et al., 2020). Longformer (Beltagy et al., 2020) uses a combination of local and global attention mechanisms to save on computational complexity and enables the processing of up to 4096 tokens. A number of other works (Dai et al., 2022; Mamakas et al., 2022) show that transformer-based architectures that can capture longer text boast major benefits, even more so when augmented with strategies like sparse-attention and hierarchical networks. This again underlines an important direction for verbose legal datasets. Our contributions can be summarized as follows:

- We conduct experiments to compare and analyze the zero-shot performance of three general LLMs to start-of-the-art in-domain models on the LEDGAR subset of LexGLUE (Chalkidis et al., 2022). We analyze our results and provide insights for further research.
- We provide an overview of the most recent LLM research, the benchmarks and datasets developed for legal NLP, the challenges faced when applying them to legal tasks, and popular approaches that try to solve them. We believe this to be a useful primer for anyone looking to get a bird's eye view of the field.

2 Related Work

In this section, we outline the relevant research on LLMs, efforts using them for legal domain tasks, and finally the benchmarks and datasets.

2.1 Large Language Models

OpenAI GPT: GPT (Generative Pre-trained Transformer) (Radford et al., 2019; Brown et al., 2020) and the popular ChatGPT variant developed by OpenAI are a family of transformer-decoder pre-trained with a vast amount of text data to perform generative and language modeling tasks and allows a reasonable context length sufficient to carry out long-document processing. For instance, GPT 3.5 allows a maximum of 4096 tokens, and GPT 4 allows a stunning maximum of 32,768, ideal for data consisting of long sequences.

Google PaLM: PaLM (Pathways Language Model) (Chowdhery et al., 2022; Anil et al., 2023) is an LLM having 540 billion parameters that was trained on the Pathways architecture. Although PaLM was initially trained to handle sequence lengths of up to 2048 tokens, it was increased to 8096 in the 340 billion parameter PaLM 2 for a longer comprehension of the input.

Meta LLaMA: LLaMA (Large Language Model Meta AI) (Touvron et al., 2023) is a collection of foundation language models ranging from 7 billion to 70 billion parameters. It was pre-trained natively on 2048 input tokens, but recent research has shown that the context length of LLMs can be extended efficiently with minimal training steps (Peng et al., 2023) and they have released two variations of LLaMA boasting a context length of 64k and 128k respectively.

TII Falcon[1]: As of the time of writing, this work by Technology Innovation Institute (TII) is not yet published but the model has been released by them. It boasts of being the largest open-source model to date of writing having 180 billion parameters, and also the highest ranking model on the Huggingface Leaderboard. It includes models with 180B, 40B, 7.5B, and 1.3B parameters, trained on TII's RefinedWeb dataset (Penedo et al., 2023).

2.2 LLMs on the legal domain

LexGPT: (Lee, 2023) finetune GPT-J models on the Pile of Law dataset (Henderson et al., 2022) and experiment with generative models for legal classification tasks. They observe that fine-tuning such out-of-the-box GPTs do not beat the state-of-the-art and in fact, provides low performance compared to discriminative models. This insightfully shows the need to bridge the gap between powerful LLMs for the legal domain.

PolicyGPT: This work (Tang et al., 2023) demonstrates how LLMs in zero-shot settings can perform remarkably well in text classification of privacy policies on several baseline LLMs. This points out how a LegalLLM may hold promise in enhancing performance on other general tasks.

Zero-and-Few-shot GPT: (Chalkidis, 2023)

[1] https://falconllm.tii.ae/falcon.html

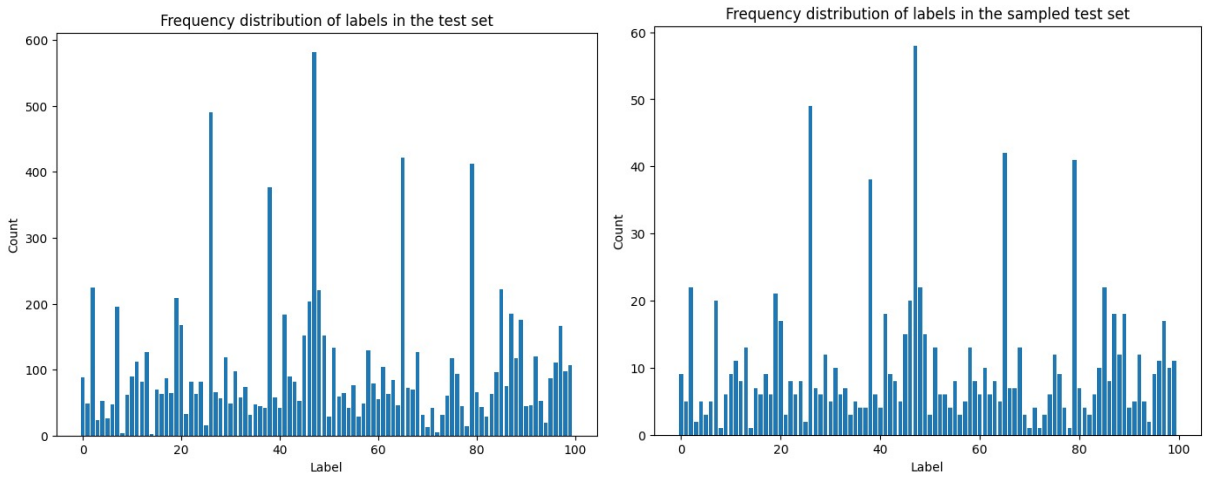

Figure 1: The frequency distributions of the 100 LEDGAR labels in the original LEDGAR test set from LexGLUE (left); and in our sampled test set of 1000 examples (right)

conduct experiments most similar to ours. This study evaluates the performance of ChatGPT on the LexGLUE benchmark in both zero-shot and few-shot settings (for the latter, examples were given in the instruction prompt, which seems to benefit the model when the number of examples and labels are around the same). They find that ChatGPT performs very well, but severely lacks in performance compared to smaller models trained on in-domain datasets.

Resonating these findings, the work of (Savelka, 2023) investigates how an LLM (a GPT model) performs on a semantic annotation task in the zero-shot setting, without being fine-tuned on in-domain datasets. The LLM is primed with a short sentence description of each annotation label and is tasked with labeling a short span of text. They observe that while the LLM performs surprisingly well given the zero-shot setting, its performance was still far off from the model that was trained on the in-domain data. In summary, both these studies highlight the potential fine-tuned LLMs can bring to the legal domain.

2.3 Datasets and Benchmarks

LexGLUE: (Chalkidis et al., 2022) present a unified evaluation framework to benchmark models. The datasets and tasks were curated from other sources of data considering various factors into account such as availability, size, difficulty, etc. They present scores of various Pre-trained Language Models (PLMs) on their benchmark. While doing so, they point out interesting results that suggest that PLMs fine-tuned on legal datasets and tasks do perform better, albeit PLMs fine-tuned on only one sub-domain don't improve on performance on the same sub-domain. Put together, their observations point out the need for a general LegalLLM (powerful enough to outperform other models on all criteria of the benchmark).

LegalBench: (Guha et al., 2023) This benchmark comprises 162 tasks representing six distinct forms of legal reasoning and outlines an empirical evaluation of 20 LLMs. They demonstrate how LegalBench supports easing communication between legal professionals and LLM developers by using the IRAC framework in the case of American law. They observe that LLMs typically perform better on classification tasks than application-based ones. They also find that for some tasks, in-context examples are not required, or only marginally improve performance. They thus conclude that the task performance in LLMs is mostly driven by the task description used in the prompt.

Pile of Law: (Henderson et al., 2022) The surge in LLM development emphasizes the need for responsible practices in filtering out biased, explicit, copyrighted, and confidential content during pre-training. Present methodologies are ad hoc and do not account for context. This paper outlines a method for filtering in the legal domain that handles the trade-offs. Thus, Pile of Law, a

growing 256GB dataset of open-source English legal and administrative data was introduced to aid in legal tasks. It allows for the understanding of government-established content filtering guidelines and illustrates various ways to learn responsible data filtering from the law.

MultiLegalPile: (Chalkidis et al., 2021) The MultiLegalPile is a 689 GB substantial dataset that spans 24 EU languages across 17 jurisdictions. It addresses the scarce availability of multilingual pre-training data in specific domains such as law, encompassing diverse legal data sources with varying licenses. With further pre-training of XLM-R models (Conneau et al., 2019), the study attains new SotA results on LEXTREME (Niklaus et al., 2023). The transformation of the XLM-R base model into a Longformer yields a fresh SotA in four LEXTREME datasets. In certain languages, monolingual models substantially outperform the XLM-R base model, achieving language-specific SotA in five languages. In LexGLUE, English models secure SotA in five of seven tasks.

3 Experimental Setup and Results

In this section, we describe our experimental approach, along with specifics of our evaluations.

3.1 Dataset and Metrics

We use the LEDGAR subset of the LexGLUE benchmark for our experiments due to its readiness to work on LLMs (for example, the other datasets do not have the label names; only the label indices are provided). Given a provision contract, the model is expected to classify the contract from 100 given labels of EDGAR themes. As mentioned, there is a high imbalance of data in datasets containing legal corpora. In particular, refer to Figure 1 for the label distribution in the LEDGAR subset of the LexGLUE benchmark. This could result in difficulties such as biased models, poor generalization, and classification scores due to data imbalance. To overcome these difficulties and enhance model evaluations, typically the F1-score is reported for such models instead of accuracy. The macro-F1 score is an even better metric in the case of data imbalance compared to micro-F1, and it is due to this reason that the macro-F1 scores are typically lower than the micro-F1 on legal tasks.

As for the sequence lengths, (Chalkidis, 2023) report the average token length of the instruction-following examples in all the LexGLUE subsets - the highest being 3.6k tokens. This restricts the capability of LLM performance due to truncation as noted earlier, and this is also highlighted in the study - few-shot settings could not be evaluated for datasets having averaged token length of more than 2k for a single example, and in many cases, the prompts were already truncated up to 4k tokens (the maximum limit of ChatGPT). The average token length of the LEDGAR subset is 0.6k.

3.2 Setup

For our experiments, we use the LEDGAR subset of the LexGLUE dataset. As baselines, we take three LLMs - ChatGPT (GPT-3.5), LLaMA-2 (70b), and Falcon (180b). As the models are very large, we use Huggingface Chat for LLaMA and Falcon. Due to this constraint, we only evaluated on a subset of 1,000 examples. However, we made sure that the subset had a label frequency distribution close to the original dataset 1 so that the evaluations may be generalized as much as possible.

We use zero-shot prompting to evaluate the above-mentioned LLMs, building on their benefit as explained earlier by other works (Tang et al., 2023; Guha et al., 2023). However, in the custom instructions (in ChatGPT) and system instructions (in Huggingface) we entered the list of EDGAR theme classes that the model should choose from. In the same fashion, to ensure that the model does not generate anything out of the list, we explicitly mentioned this as an instruction. The exact instructions that we use are provided in the appendix.

Model	**mic. F1**	**mac. F1**	**# params.**
Falcon-Chat	70.9	60.7	180b
ChatGPT	70.6	58.7	175b
LLaMA-Chat	70.4	59.6	70b
LexGPT	*83.9*	*74.0*	6b
LegalBERT	**88.2**	**83.0**	0.11b

Table 1: Comparison of general LLMs (first three models, tested on a zero-shot setting) to models fine-tuned on legal-domain datasets (last two). The current Legal-LLM is LexGPT, but the much smaller LegalBERT shows state-of-the-art performance on LEDGAR.

3.3 Results and Discussion

For our experiments, we use three baseline general-purpose chat variants - ChatGPT-3.5, Falcon-180b,

and LLaMA-2-70b - and present the results in Table 1. General-purpose LLMs perform less well than smaller in-domain models. The best general LLM, Falcon-Chat, performs 19.2% mic-F1 and 26.8% mac-F1 lower than the best in-domain model, LegalBERT, which itself is much smaller than the LexGPT, the LegalLLM. Our findings echo that of (Chalkidis, 2023).

Notably, for class labels with only one example in our sampled test set, the three chat-variants surprisingly show the same results: they fail to predict them correctly excepting the *Qualification* label (the others being *Assigns, Books, Powers, and Sanctions*. Similarly, *Indemnity* is always misclassified as *Indemnifications* (three examples in total). Further, labels that are semantically similar are difficult for the models to handle since they frequently mislabel such contract provision themes. For example, (*Taxes*, *Tax Withholdings* and *Withholdings*) is almost always labeled as *Tax Withholdings* by all the models. (*Jurisdictions, Submission To Jurisdiction, Consent To Jurisdiction*) is almost always labeled as *Submission To Jurisdiction* in the case of ChatGPT and *Jurisdiction* in the case of Falcon and LLaMA. As for (*Applicable Laws, Governing Laws, Compliance With Laws*), we observe that *Governing Laws* was easiest to predict with an average accuracy of 90%, and *Compliance With Laws* with 80%, but *Applicable Laws* performs very poorly with 0% accuracy for LLaMA and Falcon and 20% for ChatGPT - predicting only one from a total of 5 samples correctly. However, in the case of (*Payments, Fees, Interests* the models seem to predict them correctly in about 60% of the cases, with *Payments* appearing at least once for *Fees* and *Interests*. On average, only 95 of the 100 classes in the reference labels are present in the predictions.

3.4 Subjective Analysis

Our findings highlight that the perceived advantages LLMs have over BERT-based models (such as the sheer amount of large parameters, extended context length, and the amount of pre-training knowledge), cannot substitute for the obvious edge in-domain data gives to the much smaller models. Even when the LLM is trained so (LexGPT), it couldn't perform as well as the discriminative model (LegalBERT). This could be expected as the latter is more naturally suited for the benchmark's classification tasks than generative models which are prone to issues like hallucination. Our label-wise findings seem to support this too.

However, the current legal benchmarks are limited to NLU tasks. In general, it would be ideal to have a powerful LegalLLM that can perform both generative and discriminative tasks. Our findings show that there is a unique challenge in the legal domain: if we have to build a better LegalLLM, we need to find better methods to take advantage of the in-domain legal data for LLMs as simply fine-tuning it isn't enough. As the authors of LexGPT mention, reinforcement learning from human feedback could be extremely helpful in improving LexGPT, providing ways for the first LegalLLM to produce state-of-the-art results.

However, if we limit the application of legal models to NLU tasks, our findings turn optimistic. The results show that the LLMs' ability to process large context may not be necessary for classification - we hypothesize this could be because verbose legal text could turn out to have very similar semantic content, so the additional context may not be useful. This hypothesis could be echoed by findings from (Shaikh et al., 2020), who show that a careful selection of a handful of textual features in a verbose dataset is strong enough to help statistical models achieve high accuracies for binary classification.

This in fact should be good news, as it means legal practitioners can avoid having to use or train unnecessarily large or expensive models (both carbon-wise and cost-wise). Much smaller in-domain models like LegalBERT are nevertheless superior and should be used for practical applications, as also suggested by (Chalkidis, 2023)

4 Conclusion

In this work, we examine three general-purpose LLMs' zero-shot performance on a multi-class contract provision classification task using the LEDGAR dataset of LexGLUE. Our study shows that these LLMs, even though aren't explicitly trained in legal data, can still demonstrate respectable theme classification performance. The results highlight the need for better LegalLLMs adapted to the specifics of the legal industry, which has been underexplored compared to other domains. In light of this, we also present a review of related datasets and models, which we hope will help get an overview of the field.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis. 2023. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *arXiv preprint arXiv:2304.12202*.

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*, 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. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel 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)*, pages 4310–4330, 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, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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. *arXiv preprint arXiv:1911.02116*.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

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.

SiYu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-Doc: A retrospective long-document modeling transformer. 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 2914–2927, Online. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Jieh-Sheng Lee. 2023. Lexgpt 0.1: pre-trained gpt-j models with pile of law. *arXiv preprint arXiv:2306.05431*.

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*.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 130–142, Abu Dhabi, United Arab Emirates (Hybrid). 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. *arXiv preprint arXiv:2301.13126*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. *arXiv preprint arXiv:2305.04417*.

Rafe Athar Shaikh, Tirath Prasad Sahu, and Veena Anand. 2020. Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, 167:2393–2402.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, et al. 2023. Policygpt: Automated analysis of privacy policies with large language models. *arXiv preprint arXiv:2309.10238*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

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 of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

A Custom Prompt

For reproducibility, we present the prompts that we use for all our experiments. The following is the entry to the Custom Instructions setting of ChatGPT. For HuggingChat, we simply provide both the instructions to the Custom System Prompt box.

What would you like ChatGPT to know about you to provide better responses? *I want you to be an EDGAR contract provision classifier. Given a contract provision, you should correctly identify the EDGAR theme. Do not give any explanations.*

How would you like ChatGPT to respond? *One answer from the following list: [{{paste the list here}}]. Do not give an option that is not in the list.*