Evaluating Customized vs. Generalist Transformer-based Models for Legal Contract Classification

Amrita Singh, H. Suhan Karaca, Aditya Joshi, Hye-young Paik, Jiaojiao Jiang
School of Computer Science and Engineering
University of New South Wales (UNSW), Sydney

Abstract

Despite advances in legal NLP, no comprehensive evaluation of Transformer-based models customized for legal tasks (referred to as 'legal-specific' models in this paper) exists for contract classification tasks. To address this gap, we present an evaluation of 13 legal-specific transformer-based models on 3 English-language contract classification tasks and compare them with 9 generalist models. The results show that legal-specific models consistently outperform generalist models, especially on tasks requiring nuanced legal understanding. They also help reduce misclassification of rare classes in imbalanced datasets. Legal-BERT and Contracts-BERT establish new SOTAs on two of the three tasks, despite having 69% fewer parameters than the best-performing generalist models. We also identify CaseLaw-BERT and LexLM as strong additional baselines for contract classification. Our results highlight the shortcomings of generalist models, emphasizing the need for domain-specific customization, particularly in the context of legal applications.

1 Introduction

Recent work suggests that open-source legal-specific models offer a promising, cost-effective, and privacy-preserving alternative to generalist models (Singh et al., 2025; Bhambhoria et al., 2024; Chalkidis et al., 2020). However, despite their advantages, these models remain significantly underutilized in current legal contract classification tasks, as noted by Singh et al. (2025). As illustrated in Figure 1, legal-specific models are rarely evaluated, and mostly generalist models are preferred in prior work on three popular and freely available contract classification tasks: Unfair Contractual Terms Identification (Lippi et al., 2019; Chalkidis et al., 2022), Contractual Provision Topic Classification (Tuggener et al., 2020; Chalkidis et al., 2022), and Agent-Specific Deontic Modality Detection (Sancheti et al., 2022).

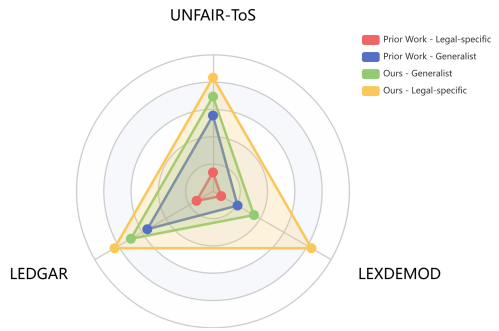

Figure 1: Comparison of our legal-specific models evaluation and coverage with prior work.

Despite the legal nature of documents/tasks, researchers have continued to favor generalist models over legal-specific models. This holds true not only for decoder models (which is the predominant context in which 'customization' is used) but also for encoder models that must be adopted to the legal domain. In some cases, legal-specific models are excluded entirely. For instance, recent studies such as Guha et al. (2023) and Singh et al. (2024), which explicitly focus on legal downstream tasks, do not include any legal-specific models in their benchmarking evaluations. Therefore, this paper addresses the Research Question **(RQ):** *How do legal-specific models perform compared to generalist models on legal contract classification tasks?* To address this question, we present a comprehensive evaluation of 13 open-source legal-specific models with 9 generalist models across the three distinct contract classification tasks. Our results reveal consistent improvements in performance for legal-specific models, particularly on tasks where legal and domain-specific semantics are critical. This benchmark serves as a resource for the community, offering a clearer understanding of model suitability

and performance across tasks and model types. The contributions of this work are as follows: (a) *To the best of our knowledge, we present the first benchmarking of multiple legal-specific models across multiple contract classification tasks*; (b) We systematically compare their performance with that of generalist models; (c) We identify model strengths, weaknesses, and task-specific challenges, offering insights for future research and deployment.

2 Contract Classification Tasks and Datasets

To evaluate the effectiveness of models customized for the legal setting vis-a-vis generalist models, we consider the following factors to select the datasets and tasks. We select the **language** to be English due to the availability of datasets and models. **Publicly available**, well-documented datasets are used, each large enough for stable training and evaluation. Proprietary, non-public, and very small datasets (under 3K sentences) are avoided to ensure reproducibility and generalizability. This criterion modifies and adapts the selection guidelines of Chalkidis et al. (2022). Datasets are chosen where SOTA generalist language models do not achieve near-perfect performance (Lippi et al., 2019; Tuggener et al., 2020; Sancheti et al., 2022), ensuring that benchmarking legal-specific language models remains **challenging**. For the chosen tasks to reflect **relevance and diversity**, we select tasks that test a model's understanding of legal language, structure, and semantics. As shown below, three distinct tasks are selected, each using a different dataset and representing a unique contract classification scenario in terms of dataset size and number of classes.

Consequently, the following datasets are selected: **UNFAIR-ToS:** The UNFAIR-ToS dataset (**train (5.5k), dev (2.3k), and test (1.6k)**) from Chalkidis et al. (2022) is a **multi-label classification** dataset that is used to identify unfair contractual terms in Terms of Service (ToS) documents from online platforms like YouTube. Each sentence is annotated with one or more of **8 unfairness categories, plus 1 unlabeled class** for sentences that do not indicate any potential violation of European consumer law EU consumer protection law.

LEDGAR: The LEDGAR dataset (**train (60k), dev (10k), and test (10k)**) from Chalkidis et al. (2022) is used to classify the principal topic of provisions in Exhibit 10 material contracts (e.g., employment, lease, non-disclosure) filed with the US Securities and Exchange Commission (SEC) via EDGAR. Each provision (paragraph) is labeled with one of **100 contract topics**, making it a **multi-class classification** task.

LEXDEMOD: The LEXDEMOD dataset (**train (4.2k), dev (330), and test (1.7k)**) from Sancheti et al. (2022) detects deontic modality in agent-based contract clauses from lease agreements sourced from the LEDGAR dataset. Each clause (sentence) is annotated with one or more of **6 deontic modality types plus 1 none class** , making it a **multi-label classification** task. Labels are linked to an agent (party) in the sentence, representing their deontic status (e.g., Obligation, Entitlement, Prohibition). *The train/dev/test split is as reported in the original paper*. Statistics and examples for all datasets are in Appendix A while the class distribution is in Figure 3.

3 Experiment Setup

We perform task-specific fine-tuning using 13 legal-specific models on three datasets: LEDGAR, UNFAIR-ToS, and LEXDEMOD. We consider ten pre-trained encoder-based legal-specific models for fine-tuning. Nine of these are base-variant encoder models: Legal-BERT (Chalkidis et al., 2020), Contracts-BERT (Chalkidis et al., 2020), LegalRoBERTa (Geng et al., 2021), CaseLaw-BERT (Zheng et al., 2021), InLegalBERT, InCaseLawBERT, and CustomInLawBERT (Paul et al., 2023), Legal-XLM-R (Niklaus et al., 2024), and LexLM (Chalkidis* et al., 2023). One large-variant model, PoL-BERT (Henderson et al., 2022), is included, as its base version is not present. We also evaluate two legal decoder-based models, AdaptLLM (Cheng et al., 2024) and SaulLM-7B (Colombo et al., 2024), along with one legal encoder-decoder model, LexT5 (T.y.s.s et al., 2024), as it is the only model of its kind among legal-specific decoder-based and encoder-decoder models. Table 5 in Appendix B summarizes the key characteristics of each model, and a detailed description of them is provided in the same appendix. A detailed experimental setup is provided in Appendix C. We compare these 13 legal-specific models with 9 generalist models. These include six base variant encoder models: BERT (Devlin et al., 2019), RoBERTa-base (Liu

	Method	Model	# Params	UNFAIR-ToS		LEDGAR		LEXDEMOD	
				μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1
Baselines reported from: (Chalkidis et al., 2022), (Sancheti et al., 2022)	Generalist Models	BERT	110M	95.6	81.3	87.6	81.8	-	75.61
		RoBERTa-base	125M	95.2	79.2	87.9	82.3	-	75.66
		DeBERTa	139M	95.5	80.3	88.2	83.1	-	-
		Longformer	149M	95.5	80.9	88.2	83.0	-	-
		BigBird	127M	95.7	81.3	87.8	82.6	-	-
		RoBERTa-large	355M	95.8	81.6	88.6	83.6	-	77.88
(Biswas et al., 2025)	Legal-KDD-Student	DistilBERT	66M	95.9	76.2	88.2	82.3	-	-
(Shin and Kaneko, 2025)	GATs	Hierarchical BERT	110M	95.6	81.5	88.4	83.1	-	-
Proposed	Generalist Models	Llama-3.2	3B	95.9	80.3	85	76.1	76.2	71.4
		Mistral	7B	96.0	80.7	86.4	79.2	76.0	71.2
	Legal-specific Models	Legal-BERT	110M	96.0	82.2	88.2	82.5	81.23	78.01
		Contracts-BERT	110M	96.2	83.4	87.9	82.2	80.17	77.71
		Legal-RoBERTa	125M	95.4	81.1	87.7	81.9	80.12	76.70
		CaseLawBERT	110M	96.1	83.2	87.6	80.9	80.32	77.75
		PoL-BERT	340M	94.6	77.9	86.0	79.1	41.35	15.75
		InLegalBERT	110M	95.6	81.7	87.9	82.0	80.21	77.89
		InCaseLawBERT	110M	95.5	81.1	87.5	82.1	79.16	76.83
		CustomInLawBERT	110M	95.5	79.9	87.7	81.8	78.16	75.35
		LexLM	124M	95.9	81.7	87.8	81.3	80.39	77.46
		Legal-XLM-R	184M	94.9	78.2	87.7	81.7	80.62	77.56
		LexT5	220M	95.4	79.8	84.9	76.1	76.50	73.30
		AdaptLLM	7B	96.5	83.2	85.2	76.8	76.2	70.0
		SaulLM	7B	96.0	81.0	86.6	79.4	76.5	72.7

Table 1: Performance of legal-specific and generalist models on three tasks: UNFAIR-ToS, LEDGAR, LEXDEMOD. Metrics: micro-F1 (μ-F1) and macro-F1 (m-F1). Blue highlights the best generalist, and red highlights the best legal-specific model performance.

Legal Specific Models	Mean ± Std	
	μ-F1	m-F1
Legal-BERT	88.48 ± 6.03	80.90 ± 2.05
Contracts-BERT	88.09 ± 6.55	81.10 ± 2.45
Legal-RoBERTa	87.74 ± 6.24	79.90 ± 2.29
CaseLawBERT	88.01 ± 6.45	80.62 ± 2.23
PoL-BERT	73.98 ± 23.34	57.58 ± 29.58
InLegalBERT	87.90 ± 6.28	80.53 ± 1.87
InCaseLawBERT	87.39 ± 6.67	80.01 ± 2.29
CustomInLawBERT	87.12 ± 7.09	79.02 ± 2.71
LexLM	88.03 ± 6.33	80.15 ± 1.91
Legal-XLM-R	87.74 ± 5.83	79.15 ± 1.82
LexT5	85.60 ± 7.73	76.40 ± 2.66
AdaptLLM	85.97 ± 8.31	76.67 ± 5.39
SaulLM	86.37 ± 7.96	77.70 ± 3.60

Table 2: Aggregated scores (Mean ± Std) across three contract classification tasks. Red, blue, and green highlights indicate the first, second, and third best performances, respectively.

et al., 2019), DeBERTa (He et al., 2021), Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), and DistilBERT (Sanh, 2019), along with one large variant, RoBERTa-large (Liu et al., 2019). We also compare two generalist decoder-based models, Llama-3.2 (Grattafiori et al., 2025) and Mistral (Jiang et al., 2023).

4 Results and Analysis

Generalist models: Table 1 reports the test results of models across all three tasks, comparing legal-specific and generalist models. Among generalist models, RoBERTa-large (355M) outperforms others, including larger decoder-based models like Llama-3.2 (3B) and Mistral (7B). Figure 3 shows that all datasets suffer from extreme long-tail class imbalance. Table 1 confirms that decoder-based models perform poorly on rare classes, as reflected in their lower macro-F1 scores, while RoBERTa-large performs consistently better. This suggests that encoder-based models are more robust to extreme class imbalance than decoder-based models. Despite having fewer parameters, RoBERTa-large achieves higher micro-F1 and macro-F1, demonstrating that *bidirectional context learning and discriminative fine-tuning are more crucial than model scale for legal classification tasks with long-tail distributions*.

Legal-specific models: Legal-specific models such as Contracts-BERT and Legal-BERT (110M) outperform RoBERTa-large on UNFAIR-ToS and LEXDEMOD, respectively, despite having 69% fewer parameters. RoBERTa-large remains the best model for LEDGAR. Still, Legal-BERT delivers equivalent performance compared to generalist base variant models on this task, suggesting that both model size and task characteristics influence performance. The larger, legal-specific encoder-based models may be better suited for LEDGAR, as the dataset is very large. *Overall,*

legal-specific base models deliver competitive performance and set new SOTAs on two of the three tasks, demonstrating the effectiveness of domain-specific pretraining, even at the base variant of models.

Generalist models vs. Legal-specific models: We further analyze the errors made by the best generalist model (RoBERTa-large) and examine whether the best legal-specific model (Contracts-BERT) corrects them on the Unfair-ToS dataset. As shown in Figure 2, RoBERTa-large majorly misclassifies clauses belonging to rare categories, particularly *Limitation of Liability* and *Unilateral Termination*, whereas Contracts-BERT correctly classifies them, suggesting that domain-specific pretraining on contract corpora (refer Table 5) enables better representation of rare legal clause types that generalist models fail to capture. This analysis is further extended to decoder-based models by comparing generalist and their corresponding legal-specific variants on the LEDGAR dataset, as shown in Figure 2. Specifically, Mistral and its legal-specific counterpart SaulLM, both sharing the same base architecture, are compared alongside Llama-3.2 and AdaptLLM, where AdaptLLM is built on Llama-1 but shares the same Llama architecture lineage. Across all cases, the results demonstrate that *domain-specific pretraining helps in reducing the misclassification of rare classes in long-tail legal text classification.*

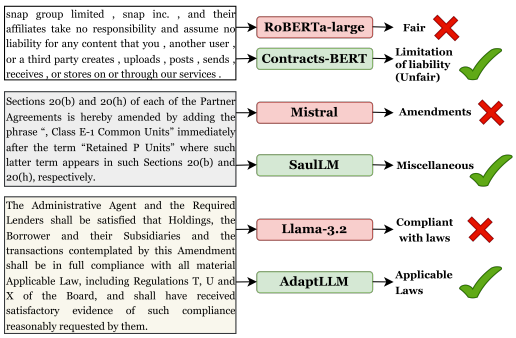

Figure 2: Example outputs showing customized models (green squares) correctly predict examples which were mis-classified by generalist models (red squares).

Legal-specific models as baselines: Table 2 presents aggregated test scores (Mean $\pm$ Std) across the three contract classification tasks. Despite class imbalance in all tasks, Legal-BERT achieves the highest aggregated μ-F1, while Contracts-BERT leads in m-F1. Across both metrics, the top positions are consistently held by four legal-specific models: Legal-BERT, Contracts-BERT, CaseLaw-BERT, and LexLM. *We conclude that these four models, Legal-BERT, Contracts-BERT, CaseLaw-BERT, and LexLM, should be considered strong baselines for contract classification tasks.*

Limitations of Recent Legal-Specific Models: Several recent legal-specific models, such as PoL-BERT, CustomInLawBERT, LexLM, Legal-XLM-R are pre-trained on large-scale legal corpora. Models like InLegalBERT and InCaseLaw-BERT are built on legal-specific base models rather than generalist models. However, older models like LegalBERT and ContractsBERT, pre-trained on just 354k and 76k legal documents (Table 5), still outperform recent base-variant legal-specific models (Table 2). This counterintuitive result can be attributed to a data distribution mismatch: more data only improves performance when it is in-distribution with the task. Recent models are trained on a broad mixture of legal genres, including court cases, legislation, and patents, which dilutes the contract-specific signal. Consequently, contracts remain underrepresented relative to other genres in these corpora. *We conclude that future legal-specific models should incorporate a more diverse and representative set of contract documents, balanced alongside other legal genres, to improve performance on contract-based downstream tasks.*

5 Conclusion

This study benchmarks 13 legal-specific and 9 generalist models across three contract classification tasks. Encoder-based models outperform larger decoder-based models in long-tail legal classification, with domain-specific pretraining consistently improving rare class recognition. Legal-specific models set new state-of-the-art results on two of the three tasks, despite having 69% fewer parameters than the best-performing generalist model. Legal-BERT, Contracts-BERT, CaseLaw-BERT, and LexLM serve as strong baselines for future research. Notably, older legal-specific models trained on smaller, more focused corpora outperform recent models trained on larger, genre-diverse corpora, highlighting that in-distribution pretraining is more critical than scale. Future work will explore retrieval-augmented approaches with legal-specific models to further enhance performance on rare and underrepresented legal clause types.

Limitations

The limited availability of contract benchmark datasets in languages other than English presents a challenge for multilingual extension. As a result, this study focuses solely on English-language contract tasks, with evaluation on non-English data left for future research. This work also concentrates on the nuances of contract language and does not assess performance on other legal text types, such as statutes, court decisions, or legal opinions. Future research should broaden this evaluation to encompass a wider range of legal genres, recognizing that no single study can fully capture the entire legal domain. Additionally, this paper focuses on domain generalization alone while selecting the legal domain.

Ethical Considerations

This study uses only publicly available datasets, LEDGAR, UNFAIR-ToS, and LEXDEMOD, all of which contain contract clauses from publicly available contract documents. LEDGAR is derived from public U.S. SEC EDGAR filings, UNFAIR-ToS from company Terms of Service, and LEXDEMOD from lease clauses sourced from LEDGAR. This research does not offer legal advice, predict individual outcomes, or automate decisions affecting rights. It focuses solely on evaluating the performance of legal-specific models to inform future tools and research. While these models can support legal professionals, they are not substitutes for legal expertise. We acknowledge potential ethical risks if outputs are misused or inaccurate. By open-sourcing our evaluations, we aim to reduce reliance on proprietary tools, promote transparency, and expand access to legal AI research and development.

References

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

Rohan Bhambhoria, Samuel Dahan, Jonathan Li, and Xiaodan Zhu. 2024. Evaluating ai for law: Bridging the gap with open-source solutions. *arXiv preprint arXiv:2404.12349*.

Momojit Biswas, Anmol Singhal, and Preethu Rose Anish. 2025. Optimized domain-specific text processing with keyword knowledge distillation (kkd). In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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, 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.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin 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*, 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.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

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.

EDGAR. Sec edgar database. https://www.sec.gov/edgar/. Accessed on 24 July 2025.

EU consumer protection law. Directive 93/13/eec on unfair terms in consumer contracts, article 3. http://data.europa.eu/eli/dir/1993/13/oj. Accessed on 24 July 2025.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Saibo Geng, Rémi Lebret, and Karl Aberer. 2021. Legal transformer models may not always help. *arXiv preprint arXiv:2109.06862*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2025. The llama 3 herd of models.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

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

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. Multilegalpile: A 689gb multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: A case study on indian law. In *Proceedings of 19th International Conference on Artificial Intelligence and Law - ICAIL 2023*.

Abhilasha Sancheti, Aparna Garimella, Balaji Vasan Srinivasan, and Rachel Rudinger. 2022. Agent-specific deontic modality detection in legal language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11563–11579.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Andrew Shin and Kunitake Kaneko. 2025. Improved understanding of legal text with graph attention networks. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, pages 364–368.

Amrita Singh, Aditya Joshi, Jiaojiao Jiang, and Hyeyoung Paik. 2025. A survey of classification tasks and approaches for legal contracts. *Springer AI Review*.

Amrita Singh, Preethu Rose Anish, Aparna Verma, Sivanthy Venkatesan, Logamurugan V, and Smita Ghaisas. 2024. A data decomposition-based hierarchical classification method for multi-label classification of contractual obligations for the purpose of their governance. *Scientific Reports*, 14(1):12755.

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.

Santosh T.y.s.s, Cornelius Weiss, and Matthias Grabmair. 2024. LexSumm and LexT5: Benchmarking and modeling legal summarization tasks in English. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 381–403, Miami, FL, USA. Association for Computational Linguistics.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *CoRR*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 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 Dataset Statistics and Illustrative Examples

This appendix provides an overview of the datasets used for benchmarking legal-specific models in Table 3 and presents labeled examples from all three datasets to aid understanding. The

Dataset	Contract Type	Task	Task Type	Train /Dev/ Test	Class
UNFAIR-ToS (Chalkidis et al., 2022)	Terms of Service (Consumer Contract)	Unfair Contractual Terms Identification	Multi-label Classification	5,532/2,275/ 1,607	9
LEDGAR (Chalkidis et al., 2022)	Exhibit-10 Material Contract	Contract Provision Topic Classification	Multi-class Classification	60,000/10,000/ 10,000	100
LEXDEMOD (Sancheti et al., 2022)	Lease Contract	Agent-Specific Deontic Modality Detection	Multi-label Classification	4,282/330/ 1,777	7

Table 3: Overview of Datasets used for Benchmarking Legal-specific Models.

class distribution of all three datasets is given in Figure 3. We provide our own rationales explaining label types, which the datasets do not explicitly include. These explanations clarify why specific labels apply to given clauses (sentences) or provisions (paragraphs), as detailed in Table 4. Legal contract classification involves longer texts than typical NLP tasks like tweets or reviews. Legal-specific Transformer models such as Legal-BERT process up to 512 sub-word tokens, but many LEDGAR paragraphs exceed this limit. Figure 4 shows that numerous LEDGAR paragraphs surpass the standard context window, requiring truncation or other methods to handle long inputs. Additionally, legal texts contain specialized terminology (legalese), increasing classification complexity compared to general text.

B Description of Legal-specific Models

Table 5 summarizes their key characteristic. The detailed description of the legal-specific models used in the experiments is given below:

Legal-BERT Legal-BERT (Chalkidis et al., 2020) is a BERT-base-uncased model (110M parameters) pre-trained on 354K English legal documents, including EU and UK legislation, US contracts, and US and EU court cases. It follows the original BERT pre-training configuration and constructs its sub-word vocabulary from scratch to better capture legal terminology.

Contracts-BERT Contracts-BERT (Chalkidis et al., 2020) is a BERT-base-uncased model (110M parameters) pre-trained on 76K US contracts. It follows the original BERT configuration and retains a custom vocabulary tailored to contract language.

Legal-RoBERTa Legal-RoBERTa (Geng et al., 2021) builds on the RoBERTa-base model (125M parameters) and continues pre-training on 4.9 GB of legal text, including patent litigation documents, US court cases, and publicly available Google Patents data.

CaseLaw-BERT CaseLaw-BERT (Zheng et al., 2021) is a BERT-base-uncased model (110M parameters) pre-trained on 3.4M US federal and state court decisions from the Harvard Case Law corpus. Although originally referred to as *Custom Legal-BERT* by (Zheng et al., 2021), it is later termed *CaseLaw-BERT* by (Chalkidis et al., 2022) to distinguish it from the earlier Legal-BERT of (Chalkidis et al., 2020), highlighting its exclusive training on harvard case law. This naming convention is now widely adopted, and we follow the same in this work.

PoL-BERT PoL-BERT (Henderson et al., 2022) is a RoBERTa-large model (340M parameters) pre-trained on the *Pile-of-Law*, a 256GB corpus comprising 10M legal and administrative documents. The dataset spans a wide range of legal domains, including US federal and state court opinions (e.g., CourtListener, SCOTUS filings), regulatory documents (e.g., Federal Register, Code of Federal Regulations, SEC and IRS guidance), legislative texts (e.g., US Bills, US Code, State Codes), and other legal document sources (e.g., ECHR, Eur-Lex, ICJ/PCIJ rulings). It also includes administrative decisions from US agencies (e.g., DOJ, OLC, BVA, NLRB, EOIR, DOL), legal contracts (e.g., EDGAR filings, Atticus contracts, CFPB agreements), educational materials (e.g., open-access casebooks, exam outlines), and publicly available community-driven legal discussions.

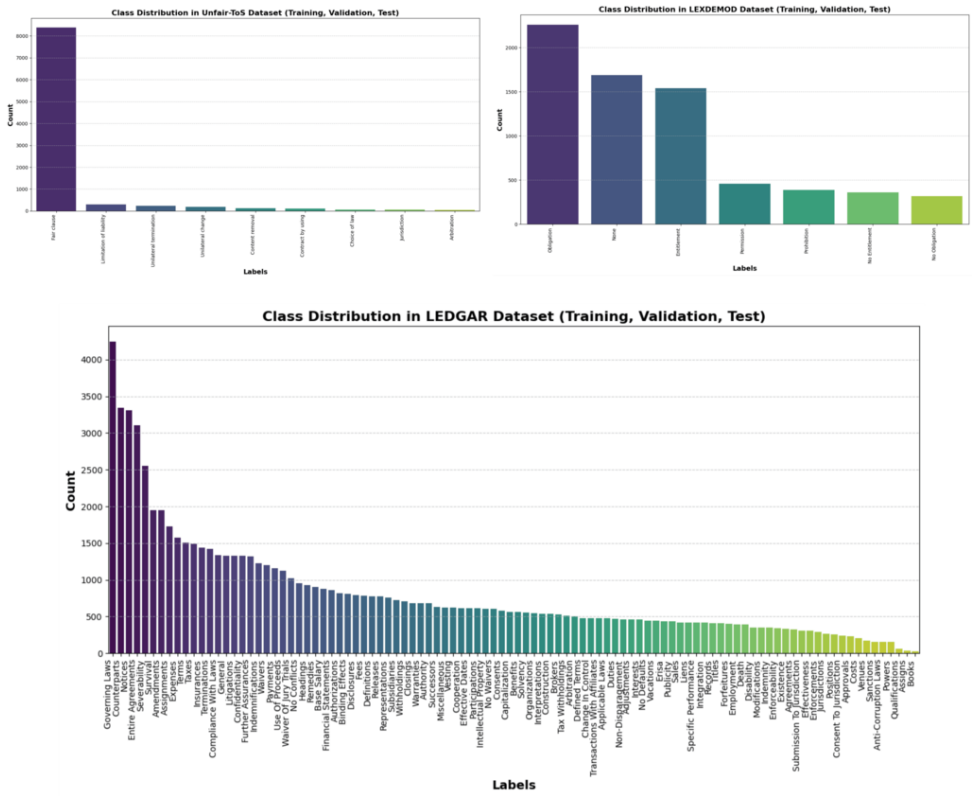

Figure 3: Class Distribution across all three datasets

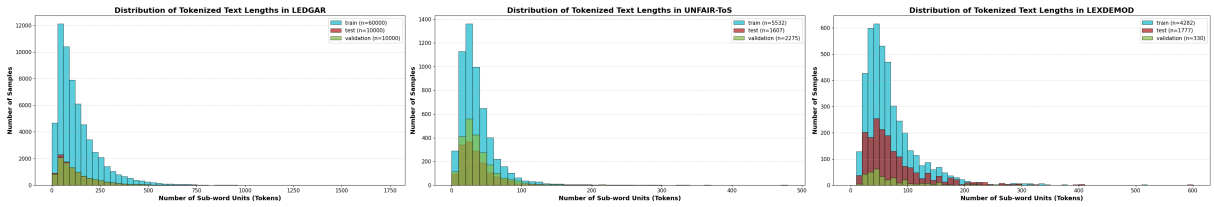

Figure 4: Distribution of text lengths, measured in Legal-BERT subword units, across all three datasets

InLegalBERT InLegalBERT (Paul et al., 2023) builds on Legal-BERT-base-uncased (Chalkidis et al., 2020), a legal-specific BERT model (110M parameters) initially pre-trained on 354K English legal documents, including EU and UK legislation, US contracts, and US and EU court cases. It is further pre-trained on 5.4M Indian legal documents, including judgments from the Supreme Court, High Courts, and District Courts, as well as Central Government Acts of India. This extended pre-training enables the model to better capture the linguistic and legal nuances of other jurisdictions, such as Indian jurisprudence.

InCaseLawBERT InCaseLawBERT (Paul et al., 2023) builds on CaseLaw-BERT-base-uncased (110M parameters), which is initially pre-trained on 3.4M US federal and state court decisions from the Harvard Case Law corpus. It undergoes further pre-training on 5.4M Indian legal documents, including judgments from the Supreme Court, High Courts, and District Courts, as well as Central Government Acts of India. This additional training enables the model to better capture the linguistic and legal nuances of other jurisdictions, particularly Indian jurisprudence.

CustomInLawBERT CustomInLawBERT (Paul et al., 2023) is a BERT-base-uncased model (110M parameters) pre-trained from scratch on 5.4M Indian legal documents, including judgments from the Supreme Court, High Courts, and District Courts, as well as Central Government Acts of India.

LexLMs LexLMs (Chalkidis* et al., 2023) include two variants: RoBERTa-base (124M parameters) and RoBERTa-large (340M parameters), both pre-trained from scratch on 5.8M legal doc-

Dataset	Example	Label	Rationale for Assigned Labels (Provided by us for better understanding)
LEDGAR	This Amendment may be executed by one or more of the parties hereto on any number of separate counterparts, and all of said counterparts taken together shall be deemed to constitute one and the same instrument. This Amendment may be delivered by facsimile or other electronic transmission of the relevant signature pages hereof.	Counterparts	This sentence states that the Amendment may be executed in multiple counterparts and that together they form a single agreement, which is a standard counterparts clause used to validate separately signed copies as one binding document.
	THIS AMENDMENT SHALL BE GOVERNED BY, AND INTERPRETED IN ACCORDANCE WITH, THE LAW OF THE STATE OF NEW YORK . The other provisions of Article IX of the Credit Agreement shall apply to this Amendment to the same extent as if fully set forth herein.	Governing Laws	This sentence specifies that New York law will govern and interpret the Amendment, which is a standard governing law clause that establishes the legal jurisdiction and framework for resolving disputes.
	Sublessee leases the Aircraft in its "as is, where is" condition. The only services, rights, or warranties to which the Sublessee is entitled to under this Sublease are those to which the Sublessor is provided under the Prime Lease.	Warranties	The sentence is labeled as warranties because it defines the rights and guarantees the Sublessee receives and limits those warranties to what the Sublessor has under the Prime Lease.
UNFAIR-ToS	Niantic further reserves the right to remove any User Content from the Service at any time and without notice and for any reason.	Content removal	This sentence is labeled as content removal unfair contractual term because it gives the provider full control to remove content at any time, for any reason, and without notice.
	amazon reserves the right to refuse service, terminate accounts , terminate your rights to use amazon services, remove or edit content , or cancel orders in its sole discretion.	Unilateral termination, Content removal	This sentence is labeled as unilateral termination and content removal because it allows Amazon to end services, remove content, or cancel orders at its sole discretion, without notice, creating an imbalance of power.
	Outside the United States and Canada. If you acquired the if you acquired the application in any other country, the laws of that country apply.	None	The sentence is labeled as none because it does not belong to any of the unfair contractual term types and is actually a fair clause.
LEXDEMOD	[lessee] Lessee will not create or permit to be created or to remain , and will promptly discharge , any lien , encumbrance or charge (including without limitation any mechanic 's , laborer 's or materialman 's lien) against the Premises or any part thereof arising from Lessee 's actions.	Prohibition, Obligation	This sentence imposes a prohibition by forbidding the Lessee (the agent) from creating or allowing liens. It also imposes an obligation by requiring the Lessee to promptly remove any such liens.
	[tenant] Tenant may, without Landlord's consent, before delinquency occurs, contest any such taxes related to the Personal Property.	Permission	This sentence grants permission to the Tenant (the agent) to contest taxes without needing the Landlord's consent, as long as it's done before delinquency.
	[landlord] Tenant shall promptly notify Landlord of any alleged defaults under the CC&Rs and/or the Oil and Gas Lease .	Entitlement	The Landlord, as the agent, holds an entitlement to receive notice from the Tenant about alleged defaults.

Table 4: Overview of all three Datasets with Examples, Labels, and Author-Provided Rationales

uments from multiple English-speaking jurisdictions. The corpus covers a wide range of sources, including EU legislation, EU and ECtHR court decisions, UK legislation and court cases, Indian court decisions, Canadian legislation and court decisions, US court decisions, US legislation, and US contracts. The dataset is designed to ensure broad jurisdictional and document-type coverage, with US legal texts comprising the largest portion. This large-scale, English legal-domain pretraining enables LexLMs to support robust legal language understanding across common law and mixed legal systems.

Legal-XLM-R Legal-XLM-R (Niklaus et al., 2024) includes two variants: RoBERTa-base (124M parameters) and RoBERTa-large (340M parameters), both pre-trained from scratch on a multilingual legal corpus comprising 59M docu-

Legal-Specific Model	Pre-training Corpora	# Doc	Base Model
Legal-BERT (Chalkidis et al., 2020)	EU Legislation, UK Legislation, European Court of Justice (ECJ) Cases, European Court of Human Right (ECHR) Cases, US Court Cases, US Contracts	354K	BERT-base-uncased
Contracts-BERT (Chalkidis et al., 2020)	US Contracts	76K	BERT-base-uncased
Legal-RoBERTa (Geng et al., 2021)	Patent Litigations, US Court Cases, Google Patents Public Data	-	RoBERTa-base
CaseLaw-BERT (Zheng et al., 2021)	Harvard Case Law (US federal and State courts)	3.4M	BERT-base-uncased
PoL-BERT (Henderson et al., 2022)	Court Opinions, Government, Publications, Contracts, Statutes, Legal Analyses, Regulations, and, more from US and EU	10M	RoBERTa-large
InLegalBERT (Paul et al., 2023)	Indian Supreme Court, High Court, and District Court Cases, Central Government Acts of India	5.4M	Legal-BERT-base-uncased
InCaseLawBERT (Paul et al., 2023)	Indian Supreme Court, High Court, and District Court Cases, Central Government Acts of India	5.4M	CaseLaw-BERT-base-uncased
CustomInLawBERT (Paul et al., 2023)	Indian Supreme Court, High Court, and District Court Cases, Central Government Acts of India	5.4M	BERT-base-uncased
LexLM (Chalkidis* et al., 2023)	EU Legislation and Case Law, UK Legislation and Case Law, Canadian Legislation and Case Law, U.S. Case Law and Contracts, ECHR Case Law, and Indian Case Law	5.8M	RoBERTa-base
Legal-XLM-R (Niklaus et al., 2024)	Different Countries Case laws and legislation, US/EU contracts, and other legal-specific documents	59M	XLM-RoBERTa-base
LexT5 (T.y.s.s et al., 2024)	EU Legislation and Case Law, UK Legislation and Case Law, Canadian Legislation and Case Law, U.S. Case Law and Contracts, ECHR Case Law, and Indian Case Law	5.8M	T5-base
AdaptLLM (Cheng et al., 2024)	US court opinions from FreeLaw project	3.6M	LLama-1
SaulLM-7B (Cheng et al., 2024)	FreeLaw, English MultiLegal Pile, EDGAR, EuroParl, GovInfo Law Stack Exchange, Australian legal corpora, EU and UK legislation, court transcripts, and USPTO filings	-	Mistral

Table 5: Key specifications of the evaluated models, including pre-training corpora (with links), document counts, and base models used.

ments. The dataset spans 24 languages and five legal text types, including legislation and case law, collected from various jurisdictions such as Germany, Switzerland, the UK, and several other countries. This large-scale, cross-lingual pre-training enables Legal-XLM-R to support legal language understanding across multilingual and multi-jurisdictional contexts.

LexT5 LexT5 (T.y.s.s et al., 2024) is a legal-oriented sequence-to-sequence model designed to address the limitations of encoder-only architectures in legal NLP. It is pre-trained on three T5 variants, T5 Small (60M parameters), T5 Base (220M), and T5 Large (770M), using the same 5.8 million legal documents employed for LexLMs (Chalkidis* et al., 2023).

AdaptLLM AdaptLLM (Cheng et al., 2024) is a LLaMA-1 model (7B parameters) pre-trained on 3.6M US court opinions from the FreeLaw project (Gao et al., 2020). It converts raw legal texts into reading comprehension tasks, such as summarization, NLI, commonsense reasoning, and text completion, using regex-based patterns, and learns in a self-supervised manner. **SaulLM-7B** SaulLM-7B (Colombo et al., 2024) is a Mistral-based model (7B parameters) pre-trained on 94B tokens of legal text. It combines data from existing sources such as FreeLaw, the English MultiLegal Pile, EDGAR, and EuroParl, along with additional datasets including GovInfo, Law Stack Exchange, Australian legal corpora, EU and UK legislation, court transcripts, and USPTO filings. The corpus is curated from both public datasets and web-scraped legal resources to ensure broad legal coverage across jurisdictions.

C Experimental Setup

We use all publicly available legal-specific pre-trained models from Hugging Face. To ensure fair comparison, we adopt the training configuration introduced by Chalkidis et al. (2022) for the LEDGAR and UNFAIR-ToS datasets: a learning rate of 3e-5 for all nine encoder-base models and 1e-5 for the encoder-large model PoL-BERT (Henderson et al., 2022), consistent with the setting used for `RoBERTa-large`. All models are trained for up to 20 epochs with a batch size of 8, using early stopping with a patience of 3 based on development set performance. For UNFAIR-ToS, we use a maximum sequence length of 128, as in Chalkidis et al. (2022). However, we disable mixed-precision training (i.e., set fp16=False)

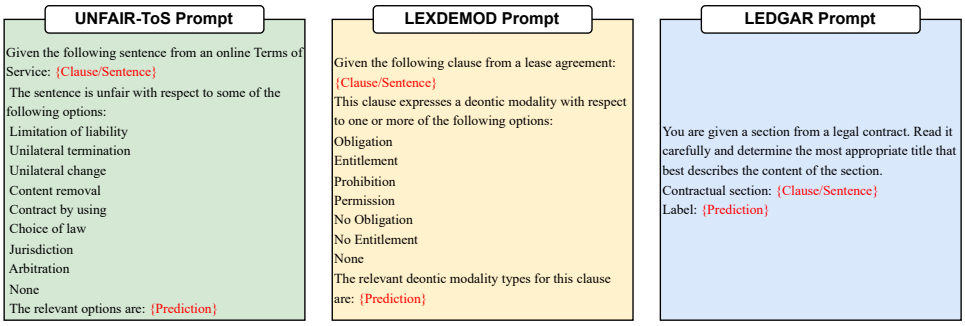

Figure 5: Prompt Templates for Instruction-Based Fine-Tuning of LexT5

to ensure stable training, which results in longer training times compared to Chalkidis et al. (2022). For LEDGAR, we reduce the maximum sequence length to 128 (from 512 used in Chalkidis et al. (2022)) to save computational resources and training time. This adjustment is necessary because, as discussed above, fp16 is disabled to ensure stable training, which leads to longer training times, and the LEDGAR dataset contains over 80k sentences, which is large. We observe only a marginal performance drop (0.1-0.4%), which we consider acceptable for efficiency. Each model is trained five times with different random seeds (1-5), and we report the test results of the best seed, following Chalkidis et al. (2022) for a fair baseline comparison. For the LEXDEMOD dataset, we follow the setup proposed by Sancheti et al. (2022), using a learning rate of 2e-5 for all encoder-based legal-specific models, including PoL-BERT, consistent with their configuration for `RoBERTa-large`. We use a batch size of 8 and apply early stopping with a patience of 3. The maximum sequence length is set to 256, as in Sancheti et al. (2022). Each model is trained five times with different random seeds, and we report the average test performance across the three best seeds, following Sancheti et al. (2022) for a fair baseline comparison. For decoder-based generalist and legal-specific models, the parameters remain the same. However, for the LEDGAR dataset, the number of epochs for models such as Mistral, AdaptLLM, and SaulLM is reduced to 3 instead of 20 due to the long training time (22-24 hours per model). This adjustment is made because fp16 is disabled for stable training, and it is necessary for LEDGAR due to its large model size and the resource constraints. In addition to encoder-decoder based legal-specific model (LexT5), we adopt instruction-based fine-tuning, which aligns better with encoder-decoder models (Wang et al., 2023). This approach pairs natural language prompts with clause inputs, enabling the model to generate the appropriate label(s) as output. The instruction templates used are listed in Figure 5. For UNFAIR-ToS, we directly use the prompt from Chalkidis (2023), originally designed for zero-shot prompting. For LEXDEMOD and LEDGAR, we design our own prompts inspired by that style. We evaluate model performance using micro-F1 (μ-F1) and macro-F1 (m-F1) to account for class imbalance. Additionally, we report the arithmetic mean with standard deviation for micro-F1 (μ-F1) and macro-F1 (m-F1) across tasks. All experiments are conducted on a single NVIDIA V100 GPU.

D Other generalist Models Considered for Exploratory Purposes

Here, we discuss the closed-source model (GPT-3.5-Turbo) results using prompting techniques, as reported by (Chalkidis, 2023). For UNFAIR-ToS, zero-shot prompting yields micro and macro F1-scores of 41.4 and 22.2, respectively. In few-shot prompting, performance improves, achieving micro and macro F1-scores of 64.7 and 32.5, though these scores remain lower than those achieved through legal-specific and generalist model fine-tuning techniques. A similar pattern occurs with LEDGAR, where zero-shot prompting results in micro and macro F1-scores of 70.1 and 56.7, respectively, while few-shot prompting produces scores of 62.1 and 51.1. Here, performance decreases with few-shot prompting, but task-specific fine-tuning using legal and generalist models remains the preferred approach. Thus, task-specific fine-tuning in the legal domain is essential for accurate legal contract classification.