Automatic Anonymization of Swiss Federal Supreme Court Rulings

Joel Niklaus[1,2,3 *] **Robin Mamié**[4 *]

Matthias Stürmer[1,2] **Daniel Brunner**[4] **Marcel Gygli**[2]

[1]University of Bern [2]Bern University of Applied Sciences
[3]Stanford University [4]Swiss Federal Supreme Court

Abstract

Releasing court decisions to the public relies on proper anonymization to protect all involved parties, where necessary. The Swiss Federal Supreme Court relies on an existing system that combines different traditional computational methods with human experts. In this work, we enhance the existing anonymization software using a large dataset annotated with entities to be anonymized. We compared BERT-based models with models pre-trained on in-domain data. Our results show that using in-domain data to pre-train the models further improves the F1-score by more than 5% compared to existing models. Our work demonstrates that combining existing anonymization methods, such as regular expressions, with machine learning can further reduce manual labor and enhance automatic suggestions.

1 Introduction

The Swiss Federal Supreme Court (SFSC) is the highest judicial authority in Switzerland. It is the final arbiter in legal disputes and ensures the uniform application of federal law throughout the country. It consists of several divisions specialized in different areas of law, including civil, criminal, administrative, and social security matters (Glaser et al., 2021). In a year, the SFSC roughly handles 7K cases and publishes its rulings. In this process, personal information must be anonymized from the rulings in order to protect involved parties. In the traditional setting, court rulings are anonymized by skilled experts. This task is highly complex, as the removal/anonymization of a word is dependent on the context it is written in. For example, *Zuerich* needs to be removed if it is part of the name of the legal entity "Zurich Insurance Group", but not if it is a reference to the city. At the SFSC, experts are already supported in their work through an application called *Anom2* (see Figure 1). *Anom2* provides access to various methods and algorithms for finding and replacing text entities (e.g., with regular expressions. The aim of this work is to enhance the capabilities of *Anom2* with Machine Learning capabilities that provide the user with more suggestions that need to be anonymized. Our results show that this approach allows users to find more elements that require anonymization.

[*] Equal contribution.

2 Related Work

For identifying elements that might require anonymization, a process called Named Entity Recognition (NER) is employed. Traditionally, NER recognizes and categorizes text parts according to a set of semantic categories like *Location (LOC)*, *Organization (ORG)*, or *Person (PER)* (Benikova et al., 2014). As these classes are not enough for the anonymization of court cases (Leitner et al., 2020) suggested expanding this list to seven coarse and 19 fine-grained classes, including entities such as *Judge (RR)*, or *Lawyer (AN)*. Using this dataset (Darji et al., 2023) fine-tuned GermanBERT (Chan et al., 2020), clearly outperforming a BiLSTM-CRF+ model. Similar approaches have been applied and tested in other languages, such as Romanian (Pais et al., 2021), Greek (Angelidis et al., 2018), Portuguese (Luz de Araujo et al., 2018), and multilingually (de Gibert et al., 2022; Niklaus et al., 2023a).

Domain-specific pretraining has flourished in the legal domain recently. Chalkidis et al. (2020) pretrained LegalBERT on EU and UK legislation, ECHR and US cases, and US contracts. Zheng et al. (2021) pretrained CaseHoldBERT on US case law, while Henderson et al. (2022) trained PoL-BERT on the 256 GB Pile of Law corpus. Niklaus and Giofré (2022) pretrained Longformer (Beltagy et al., 2020) models using the Replaced

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 159–165
December 7, 2023 ©2023 Association for Computational Linguistics

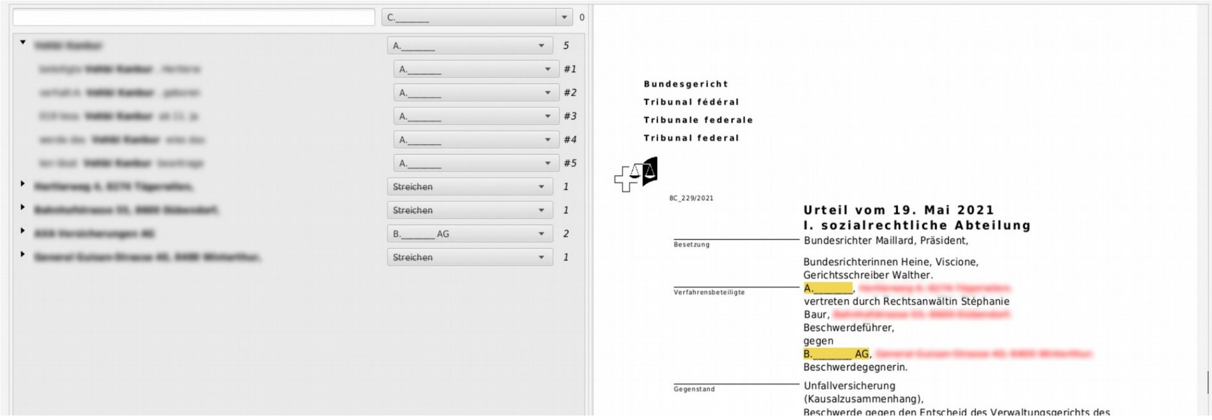


Figure 1: Main window of *Anom2*. Anonymizations are configured on the left, and the anonymized court ruling appears on the right. The system highlights completed anonymizations in gold and the current setting in yellow.

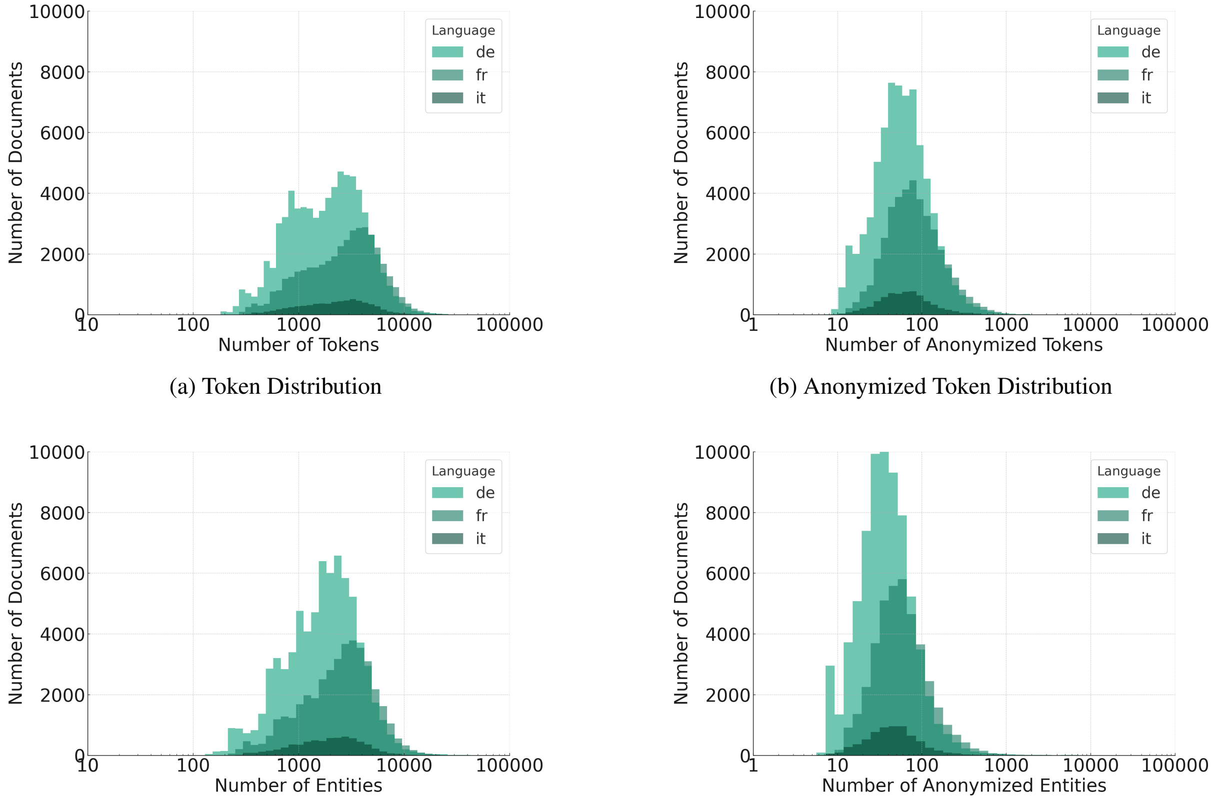
Token Detection (RTD) task on the Pile of Law. Hua et al. (2022) used RTD to pretrain Reformer (Kitaev et al., 2020) models on 6 GB of US case law. Finally, Niklaus et al. (2023b) released a large multilingual legal corpus and trained various legal models. We continue pretraining the German, French, and Italian models for 800K and 300K steps more for base and large models, respectively. Rasiah et al. (2023) pretrain models on Swiss legal data, termed Legal-Swiss-RoBERTa.

Document anonymization has a long tradition in the medical domain, where personal data need to be removed from documents. Initially, this task was handled using methods such as semantic lexicons (Ruch et al., 2000) or regular expressions to replace text occurrences. Recently, this has been expanded to include BERT-style models as well (Mao and Liu, 2019). In the legal domain, Glaser et al. (2021) worked on 1400 anonymized German rulings. Using already anonymized rulings, they trained different Recurrent Neural Networks (RNN) using BERT embeddings. Using this approach, they achieved a maximum of 68.9% precision and 79.1% recall rates. Garat and Wonsever (2022) performed similar work on 80K documents from Uruguayan courts. Our work specifically tackles court decisions by the SFSC. We compare the generic cased mBERT model (Devlin et al., 2019a) with models pre-trained on in-domain data (such as Legal-Swiss-RoBERTa-base (Rasiah et al., 2023)). We also investigate monolingual model performance in the three languages of the SFSC rulings: German, French, and Italian.

Much prior work used SFSC cases as data for their research because of wide availability in three languages, giving good coverage of the most important Swiss case law. Niklaus et al. (2021, 2022) introduced and studied judgment prediction on SFSC rulings. Brugger et al. (2023) investigated and improved multilingual sentence boundary detection in the legal domain using SFSC decisions. Christen et al. (2023) studied negation scope resolution and Nyffenegger et al. (2023) investigated how easily LLMs can re-identify persons occurring in anonymized SFSC decisions. Rasiah et al. (2023) created a large benchmark of ten text classification tasks, two text generation tasks, an information retrieval, and a citation extraction task.

3 Dataset

We used 119156 rulings (77262 German, 40099 French, 6795 Italian) Supreme court decisions and split them into sentences using Spacy (Honnibal et al., 2020). We prepared the decisions for NER based on the manual labels from the paralegals who performed manual anonymizations. The histograms in Figure 2 illustrate the distribution of four key measures, namely, Number of Tokens, Anonymized Tokens, Entities, and Anonymized Entities, in three languages: German (de), French (fr), and Italian (it). Different color schemes for each language enhance the visual interpretability of the plots. Measures concerning tokens and entities exhibit a long-tailed distribution, signifying a concentration of instances at the lower end of the value spectrum. Specifically, the distribution of Number of Tokens and Number of Entities is examined within a 10 to 100,000 range, capturing their broad spread. In contrast, anonymized tokens and entities are evaluated within a 1 to 10,000 range, reflecting their constrained distribution.



(a) Token Distribution

(b) Anonymized Token Distribution

(c) Entity Distribution

(d) Anonymized Entity Distribution

Figure 2: Histograms illustrating the distribution of (anonymized) tokens and entities across the three languages.

4 Legal Pretraining

To improve the SFSC anonymization system, we pretrained legal-specific models on diverse legal text in German, French, and Italian.
(a) We warm-start (initialize) our models from the original XLM-R checkpoints (base or large) of Conneau and Lample (2019). Model recycling is a standard process followed by many (Wei et al., 2021; Ouyang et al., 2022) to benefit from starting from an available "well-trained" PLM, rather from scratch (random). XLM-R was trained on 2.5 TB of cleaned CommonCrawl data in 100 languages.
(b) We train a new tokenizer of 32K BPEs on the training subsets to better cover legal language. However, we reuse the original XLM-R embeddings for all lexically overlapping tokens (Pfeiffer et al., 2021), i.e., we warm-start word embeddings for tokens that already exist in the original XLM-R vocabulary, and use random ones for the rest.
(c) We continue pretraining our monolingual models on our pretraining corpus with batches of 512 samples for an additional 1M/500K steps for the base/large model. We do initial warm-up steps for the first 5% of the total training steps with a linearly increasing learning rate up to $1e{-}4$, and then follow a cosine decay scheduling, following recent trends. For half of the warm-up phase (2.5%), the Transformer encoder is frozen, and only the embeddings, shared between input and output (MLM), are updated. We also use an increased 20/30% masking rate for base/large models respectively, where also 100% of the predictions are based on masked tokens, compared to Devlin et al. (2019b)[1], based on the findings of Wettig et al. (2023).
(d) We consider mixed cased models, i.e., both upper- and lowercase letters covered, similar to recently developed large PLMs (Conneau and Lample, 2019; Raffel et al., 2020; Brown et al., 2020).
(e) This leaves us with two models for each language (base and large). Additionally, we consider the multilingual legal models pretrained by Niklaus et al. (2023b) and the Swiss legal models pretrained by Rasiah et al. (2023).

5 Anonymization System

The SFSC employs an anonymization system, *Anom2*, to assist paralegals in anonymizing rul-

[1]Devlin et al. (2019b) – and much follow-up work – used a 15% masking ratio, and a recipe of 80/10/10% of predictions made across masked/randomly-replaced/original tokens.

Table 1: Evaluation Results. Best results per setup are in **bold**.

Model	Normal			Uniformizing		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Multilingual Models						
bert-base-multilingual-cased	90.72	83.76	87.10	85.85	**94.95**	**90.17**
Legal-XLM-RoBERTa-base	**94.84**	81.98	87.94	**89.93**	86.85	88.36
Legal-Swiss-RoBERTa-base	92.26	**92.57**	**92.42**	83.13	94.85	88.60
Monolingual Models						
bert-base-german-cased	95.14	80.00	86.92	**91.49**	**85.86**	**88.58**
Legal-German-RoBERTa-base	**95.40**	**80.09**	**87.07**	89.20	84.97	87.03
dbmdz/bert-base-french-europeana-cased	**95.86**	81.84	88.30	**88.92**	89.14	**89.03**
Legal-French-RoBERTa-base	95.45	**83.48**	**89.06**	88.77	**89.17**	88.97
dbmdz/bert-base-italian-cased	93.49	80.21	86.35	76.71	83.85	80.12
Legal-Italian-RoBERTa-base	**94.16**	**80.59**	**86.85**	**84.03**	**84.06**	**84.05**

ings for public access. The main UI is shown in Figure 1. Upon loading a ruling, the application auto-identifies terms requiring anonymization and lists them on the left, along with replacement text. The search function allows direct term marking for anonymization. *Anom2* uses different algorithms for the search for text that needs to be anonymized:
Conventional is based on a statistical analysis of the loaded ruling. Using *polyglot*[2] an initial set of named entities is detected. Using the specific knowledge of the format, the rubrum is dynamically detected, allowing for the labelling of important names and addresses.
BERT performs the recognition of entities to be anonymized using a BERT (Devlin et al., 2019a) model fine-tuned for NER. Entity recognition is performed on the sentence level, as the rulings are often too long for the model. This approach could lead to inconsistencies in recognition, as a term identified in one sentence might not be identified in another. This is solved in post-processing, where any identified term is automatically anonymized in the whole document.
Legal-Swiss-RoBERTa-base works analogously to the BERT method, but uses a fine-tuned *Legal-Swiss-RoBERTa-base* (Rasiah et al., 2023) model.

6 Experimental Setup

We used the following hyperparameters for all evaluated models: batch size of 64, learning rate of 5e-5, and weight decay of 0.01. We employed the seqeval metric for evaluation. We set the maximum sequence length to 192 tokens, which we determined to be the optimal trade-off between average sentence size and training time for computational efficiency. We used early stopping based on the F1-score of the validation set, which constitutes 10% of the entire dataset, following an 80-10-10 split for the training, validation, and test sets, respectively. Training ceases once the F1-score on the validation set starts to decline. Due to resource constraints (we only had two Tesla T4 GPUs) we could only run one random seed per model.

We define and configure two special parameters:
1) *TruncationStrideRatio*: We set this parameter to 0.5. When a sentence exceeds 192 tokens, we truncate it using a specific overlap strategy. The overlap consists of half of the previous snippet and half of the next snippet.
2) *NonAnonymizedSentencesRatioToAnonymizedSentences*: We set the ratio at 1.5, including only 150% of sentences without anonymization examples compared to those with examples. This minimizes data redundancy and maximizes utility.

7 Results

Table 1 presents a comprehensive evaluation of various BERT and RoBERTa-based models on two different conditions: Normal and Uniformizing. For the Normal condition, in the multilingual setting, Legal-XLM-RoBERTa-base exhibits the highest Precision at 94.84%, while Legal-Swiss-RoBERTa-base demonstrates superior Recall and F1-Score values, achieving 92.57% and 92.42% respectively. With Uniformizing, we describe the process of forcing the model to replace all occurrences of a detected term across the whole docu-

[2]See: https://polyglot.readthedocs.io

ment. This approach leads to better Recall, but reduces Precision. In the Uniformized case, again Legal-XLM-RoBERTa-base shows highest Precision, while mBERT achieves highest Recall and F1-Score. The improved Recall and F1-Score in the Normal condition show that pre-training on legal data can improve the performance of models. We observe similar behavior for the monolingual models. All models pre-trained on legal data achieve a higher F1-Score than generic monolingual models.

8 Discussion

We pretrained models on Swiss legal data and performed a detailed comparison of legal and generic models, both multilingually and monolingually in the ruling anonymization task. Our experiments indicate that pretraining on legal data improves the performance of models significantly compared to generic multi- or monolingual models.

To reduce errors in sentence splitting, we suggest future work to use legal specific sentence splitters (Brugger et al., 2023). Due to computational constraints we only experimented with base size encoder models. Future work may expand this by also testing larger models.

Acknowledgements

We greatly appreciate Google's generous support of TPUs v3-8 machines for pretraining the models. This work has been partially supported by the Swiss National Research Programme "Digital Transformation" (NRP-77) grant number 187477.

References

I. Angelidis, Ilias Chalkidis, and M. Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *JURIX*.

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

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Tobias Brugger, Matthias Stürmer, and Joel Niklaus. 2023. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset. ArXiv:2305.01211 [cs].

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.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. *arXiv:2010.10906 [cs]*. ArXiv: 2010.10906.

Ramona Christen, Anastassia Shaitarova, Matthias Stürmer, and Joel Niklaus. 2023. Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. German BERT Model for Legal Named Entity Recognition:. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 723–728, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Ona de Gibert, A García-Pablos, Montse Cuadros, and Maite Melero. 2022. Spanish datasets for sensitive entity detection in the legal domain. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22), Marseille, France, june. European Language Resource Association (ELRA)*. Dataset URL: https://tinyurl.com/mv65cp66.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Diego Garat and Dina Wonsever. 2022. Automatic Curation of Court Documents: Anonymizing Personal Data. *Information*, 13(1):27.

Ingo Glaser, Tom Schamberger, and Florian Matthes. 2021. Anonymization of german legal court rulings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 205–209, São Paulo Brazil. ACM.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. ArXiv:2207.00220 [cs].

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. 2022. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension. ArXiv:2212.08204 [cs].

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. *arXiv:2001.04451 [cs, stat]*. ArXiv: 2001.04451.

Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. 2020. A Dataset of German Legal Documents for Named Entity Recognition. *arXiv:2003.13016 [cs]*. ArXiv: 2003.13016.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 313–323, Cham. Springer International Publishing.

Jihang Mao and Wanli Liu. 2019. Hadoken: a BERT-CRF Model for Medical Document Anonymization. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, pages 720–726. CEUR-WS.org.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? ArXiv:2211.17135 [cs].

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

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023b. MultiLegalPile: A 689GB Multilingual Legal Corpus. ArXiv:2306.02069 [cs].

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2023. Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models. ArXiv:2308.11103 [cs].

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

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E. Ho, and Joel Niklaus. 2023. SCALE: Scaling up the Complexity for Advanced Language Model Evaluation. ArXiv:2306.09237 [cs].

P. Ruch, R. H. Baud, A. M. Rassinoux, P. Bouillon, and G. Robert. 2000. Medical document anonymization with a semantic lexicon. *Proceedings. AMIA Symposium*, pages 729–733.

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

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

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. *arXiv:2104.08671 [cs]*. ArXiv: 2104.08671 version: 3.