From Citations to Criticality:
Predicting Legal Decision Influence in the Multilingual Swiss Jurisprudence

Ronja Stern [1*] **Ken Kawamura** [5*]

Matthias Stürmer [1,2] **Ilias Chalkidis** [4] **Joel Niklaus** [1,2,3*]

[1]University of Bern [2]Bern University of Applied Sciences
[3]Stanford University [4]University of Copenhagen [5]Independent Scholar

Abstract

Many court systems are overwhelmed all over the world, leading to huge backlogs of pending cases. Effective triage systems, like those in emergency rooms, could ensure proper prioritization of open cases, optimizing time and resource allocation in the court system. In this work, we introduce the Criticality Prediction dataset, a novel resource for evaluating case prioritization. Our dataset features a two-tier labeling system: (1) the binary LD-Label, identifying cases published as Leading Decisions (LD), and (2) the more granular Citation-Label, ranking cases by their citation frequency and recency, allowing for a more nuanced evaluation. Unlike existing approaches that rely on resource-intensive manual annotations, we algorithmically derive labels leading to a much larger dataset than otherwise possible. We evaluate several multilingual models, including both smaller fine-tuned models and large language models in a zero-shot setting. Our results show that the fine-tuned models consistently outperform their larger counterparts, thanks to our large training set. Our results highlight that for highly domain-specific tasks like ours, large training sets are still valuable.

1 Introduction

Predicting the impact of legal cases is a critical task in the legal domain, as it aids professionals in the judicial system prioritize and manage large volumes of case law more effectively, ensuring that critical cases receive the necessary attention. Despite its significance, the task of predicting the case criticality remains relatively under-explored. Existing approaches to evaluating the importance of legal cases are primarily manual, very resource-intensive and subject to the judgments of individual annotators. This paper introduces a novel dataset , licensed under CC BY 4.0,[1] and a more challenging evaluation

---

* Equal contribution.
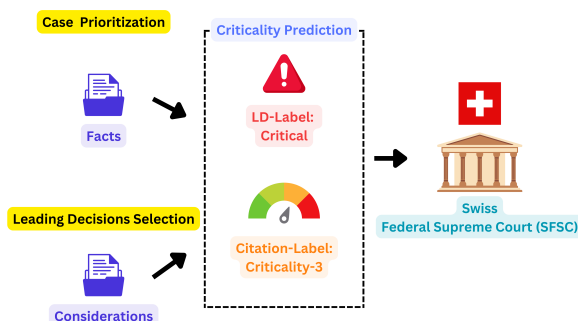[1]https://huggingface.co/datasets/rcds/swiss_criticality_prediction

Figure 1: Overview of the Criticality Prediction Task.

framework—**Criticality Prediction**—designed to predict the potential influence of Swiss Federal Court cases on future jurisprudence.

While prior work such as the Importance Prediction task for European Court of Human Rights (ECtHR) (Chalkidis et al., 2019) cases focused on predicting importance on a defined scale using human-assigned labels, our approach employs algorithmically derived labels to evaluate case criticality. Our dataset introduces a two-tier labeling system: the **LD-Label**, a binary indicator of whether a case is published as a Leading Decision (LD), and the **Citation-Label**, a more nuanced categorization based on the frequency and recency-weighted citation counts of these decisions across subsequent cases. This distinct formulation of "criticality" not only distinguishes critical cases from non-critical ones but also ranks them by their relative importance over time. The complexity of this dataset challenges even recent LLMs such as ChatGPT.

Our contributions are threefold: (1) We propose a novel Criticality Prediction task that provides a more comprehensive and challenging evaluation of case law importance. (2) We release the datasets to the community, providing valuable resources for further research in legal NLP. (3) We evaluate several multilingual models of various sizes, including fine-tuned variants to establish baselines.

897

2 Related Work

One of the most common text classification tasks in the legal domain is Legal Judgment Prediction (LJP), which involves predicting outcome of a case based on its facts. Researchers have leveraged datasets with unique characteristics to analyze and predict case outcomes across various languages, jurisdictions, and input types (Feng et al., 2022; Aletras et al., 2016; Şulea et al., 2017; Medvedeva et al., 2018; Chalkidis et al., 2019; Niklaus et al., 2021, 2022; Semo et al., 2022). While LJP is focused on the outcome of individual cases, Importance Prediction shifts the focus toward assessing the broader significance of a case.

Chalkidis et al. (2019) introduced the Importance Prediction task using cases from the European Court of Human Rights (ECtHR). In this task, ECtHR provided scores that denote each case's 'importance' to the common law. These scores, ranging from 1 (key case) to 4 (unimportant), were designed to help legal practitioners identify cases that play a crucial role in shaping jurisprudence. The labels reflect the long-term impact of a case on future rulings and the evolution of legal precedent. While the task is invaluable for identifying landmark cases, it relies on legal experts to assign the labels, making the process resource-intensive and potentially subject to subjective interpretations. Additionally, the ECtHR dataset contains only English samples, whereas our Swiss dataset is multilingual. To our knowledge, no other study has specifically addressed the task of Criticality Prediction.

3 Task and Dataset

3.1 Criticality

Understanding the legal framework of the Swiss Federal Supreme Court (SFSC) is key to defining criticality in Swiss case law. The SFSC shapes the legal landscape through its rulings, with a subset known as Leading Decisions (LDs) published separately due to their influence on the interpretation of law. We quantify case criticality using two labels: the **LD-Label** and the **Citation-Label**.

The **LD-Label** is binary, categorizing cases as *critical* or *non-critical*. SFSC cases are labeled *critical* if also published as LD, reflecting their recognized importance within the Swiss legal system. We used regular expressions to extract SFSC cases published as LD. SFSC cases not published as LD are labeled *non-critical*.
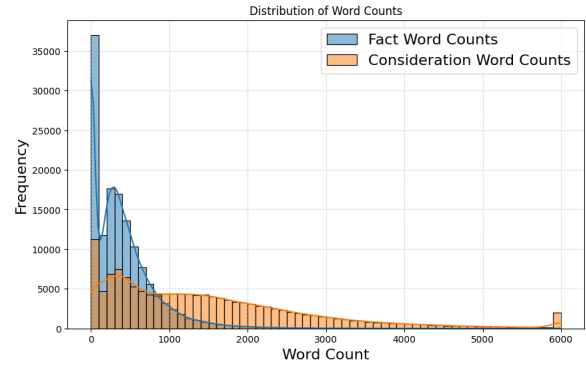

Figure 2: Facts (Blue) and consideration (Orange) length distribution measured in Words with Spacy (Honnibal et al., 2020). Those with more than 6000 words are binned together.

We developed the **Citation-Label** for a more granular measure. It counts how often each LD case is cited in SFSC cases, with less weight on older cases to prioritize recency. The score is calculated as: $score = count \times \frac{year - 2002 + 1}{2023 - 2002 + 1}$. The *count* is the citation frequency, and the weighting reduces older cases' influence. This score ranks LD cases, and we categorized them into four criticality levels—*critical-1* (least critical) to *critical-4*—based on the 25th, 50th, and 75th percentiles. More details on the constants are in Appendix A.

Unlike a prior approach (Chalkidis et al., 2019), which also used four categories, we explicitly incorporate temporal weighting to account for both the influence of a case and how its criticality shifts over time. Furthermore, our framework allows for the dynamic recalculation of scores and re-labeling of criticality as case law evolves. This adaptability ensures that our dataset reflects the ongoing changes in the legal system, whereas prior studies would require manual re-annotation as the law develops.

3.2 Criticality Prediction Task and Dataset

The Criticality Prediction (CP) task leverages two primary inputs from the Swiss Federal Supreme Court (SFCS) cases: **facts** and **considerations**. Facts describe a factual account of the events of each case and form the basis for the considerations of the court. Considerations reflect the formal legal reasoning, citing laws and other influential rulings, and forming the basis for the final ruling.

We see two distinct applications for these inputs in the Criticality Prediction task, illustrated in Figure 1. In the *Case Prioritization* task, only the facts are used as input. This produces a score indicating how critical or important a case is. The goal is to help prioritize cases, which could assist in

Table 1: Task Configurations. Label names are *Critical* (C), *Non-critical* (NC), *Critical-1* (C1) to *Critical-4* (C4).

Task Name	*Train*					*Validation*					*Test*				
	Size	Label Distribution				Size	Label Distribution				Size	Label Distribution			
		C	NC				C	NC				C	NC		
LD-Facts	**74799**	2542	72257	-	-	**12019**	580	11439	-	-	**26239**	950	25289	-	-
LD-Considerations	**87555**	2544	85011	-	-	**13386**	580	12806	-	-	**29486**	948	28538	-	-
		C-1	C-2	C-3	C-4		C-1	C-2	C-3	C-4		C-1	C-2	C-3	C-4
Citation–Facts	**2506**	782	626	585	513	**563**	186	152	131	94	**725**	137	177	224	187
Citation-Considerations	**2509**	779	624	586	520	**563**	186	154	131	92	**723**	137	177	224	185

determining which cases should be heard sooner or assigned to more experienced judges. Many court systems are overwhelmed all over the world leading to huge backlogs of pending cases. Effective triage systems, like those in emergency rooms, could ensure proper prioritization of open cases.

In the *Leading Decisions Selection* task, considerations are used for a post-hoc analysis, comparing the ruling to prior case law to assess its potential impact on future jurisprudence. This analysis is exactly what the Supreme Court does at the end of the year to arrive at the selection of leading decisions.

Our approach, by offering facts and considerations as inputs, reflects different stages of legal processing. Both tasks can utilize the LD-Label or Citation-Label, providing varying levels of granularity. This dual approach enhances the dataset's practical utility, supporting both early-stage prioritization and post-judgment evaluation, thereby addressing multiple aspects of legal workflows.

Our dataset spans from 2002 to 2023, and is partitioned into train (2002-2015), dev (2016-2017), and test (2018-2022) sets, as outlined in Table 1. We allocated a relatively large test set to accommodate longitudinal studies, including the COVID-19 pandemic years. This setup ensures that the test set represents the most recent data for realistic evaluation (Søgaard et al., 2021). The dataset consists of 138,531 total cases, with 85,167 in German, 45,451 in French, and 7,913 in Italian. See Figure 2 for length distributions of facts and considerations and Appendix Table 4 for general dataset metadata.

4 Methods

4.1 Models

We evaluated the following models: XLM-R (Base and Large) (Conneau and Lample, 2019), MiniLM (Wang et al., 2020), DistilmBERT (Sanh, 2019), mDeBERTa-v3 (He et al., 2021), X-MOD (Base) (Pfeiffer et al., 2022), SwissBERT (Vamvas et al., 2023), mT5 (Small and Base) (Xue et al., 2021), BLOOM (560M) (Scao et al., 2022), Legal-

Swiss-RoBERTa and Legal Swiss Longformer (Base) (Rasiah et al., 2023), GPT-3.5 (Brown et al., 2020), and LLaMA-2 (Touvron et al., 2023).

We fine-tuned all models per task, using early stopping on the dev set. Due to resource constraints, further fine-tuning of GPT-3.5 and LLaMA-2 is reserved for future work, with their current performance serving as baseline results. SwissBERT and Legal-Swiss models were chosen for their Swiss-specific pretraining, while the other models were chosen for their multilingual capabilities, essential for our multilingual dataset.

We evaluated GPT-3.5 and LLaMA-2 in a zero-shot setting following Chalkidis (2023). Samples were randomly selected from the validation set to prevent test set leakage for future evaluations especially for a closed model (GPT-3.5). To manage costs, we limited the validation set to 1000 samples. Our experiments focused on zero-shot classification due to the long input lengths. The prediction labels were determined using regular expressions to match model outputs to class labels. We show the prompts used in Appendix Figure 4 and Figure 5.

4.2 Metrics

We adopt the LEXTREME benchmark setup (Niklaus et al., 2023), and use aggregation of macro-averaged F1 scores with the harmonic mean to emphasize lower scores, promoting fairness across languages and input types. Scores are averaged over random seeds, languages (de, fr, it), and input types (facts or considerations), encouraging consistent performance across configurations.

5 Results

We present results in Table 3, with standard deviations in Appendix Table 5 and scores on the validation dataset in Appendix Table 6. The best performance was achieved by XLM-R$_{Large}$, with an aggregate (Agg.) score of 37.1. SwissBERT also demonstrated competitive results, with an Agg. score of 34.8. Interestingly, larger models did not always

Table 2: Configuration aggregate scores. The macro-F1 scores from the language-specific subsets of the test set are provided.

Model	LD-F	LD-C	C-F	C-C	Agg.
Languages	de / fr / it	de / fr / it	de / fr / it	de / fr / it	de / fr / it
MiniLM	57.5 / 53.9 / 52.9	68.1 / 65.4 / 64.2	12.1 / 13.1 / 6.8	24.6 / 21.9 / 17.3	25.7 / 25.7 / 16.7
DistilmBERT	56.3 / 55.6 / 56.8	67.8 / 63.9 / 64.7	20.2 / 18.2 / 20.7	22.6 / 21.6 / 22.2	31.7 / 29.7 / 31.6
mDeBERTa-v3	57.6 / 55.1 / 52.7	73.9 / 68.1 / 67.7	25.4 / 22.8 / 16.8	22.1 / 21.6 / 12.6	34.6 / 32.5 / 23.2
XLM-R$_{Base}$	59.4 / 56.3 / 56.0	70.2 / 65.4 / 62.5	20.0 / 20.6 / 23.5	26.5 / 22.1 / 23.1	33.7 / 31.5 / 33.4
XLM-R$_{Large}$	58.4 / 56.8 / 54.1	70.5 / 67.3 / 66.0	22.5 / 19.7 / 36.2	26.7 / 28.2 / 33.0	46.9 / 33.7 / 43.7
X-MOD$_{Base}$	59.0 / 56.2 / 54.8	71.1 / 68.7 / 64.1	19.8 / 17.2 / 24.4	23.2 / 24.2 / 16.4	32.1 / 30.3 / 29.5
SwissBERT$_{(xlm-vocab)}$	57.6 / 55.9 / 57.3	72.4 / 69.3 / 61.2	23.8 / 20.3 / 39.4	28.5 / 24.0 / 18.7	36.9 / 32.3 / 35.5
mT5$_{Small}$	54.8 / 51.7 / 50.3	69.2 / 61.9 / 56.4	14.2 / 16.2 / 10.5	15.9 / 18.1 / 20.2	24.1 / 26.2 / 21.9
mT5$_{Base}$	54.1 / 52.1 / 50.3	66.4 / 61.9 / 56.8	10.6 / 16.3 / 16.9	18.7 / 18.7 / 22.1	22.1 / 26.6 / 28.2
BLOOM-560m	55.1 / 53.2 / 50.9	64.6 / 65.3 / 56.2	12.6 / 16.1 / 7.1	9.5 / 13.6 / 5.1	18.3 / 23.6 / 10.7
Legal-Swiss-R$_{Base}$	59.3 / 58.4 / 55.5	73.8 / 69.4 / 68.6	24.3 / 20.5 / 10.5	26.2 / 25.3 / 14.0	36.5 / 33.4, 20.1
Legal-Swiss-R$_{Large}$	58.3 / 55.7 / 53.6	71.9 / 68.5 / 66.7	23.0 / 21.3 / 38.7	28.5 / 26.0 / 9.0	36.5 / 33.9 / 23.4
Legal-Swiss-LF$_{Base}$	60.7 / 58.3 / 55.5	74.8 / 70.0 / 67.8	25.3 / 21.5 / 18.4	29.2 / 26.7 / 9.9	38.6 / 34.7 / 21.3

Table 3: Results using Macro F1, with highest values in **bold**. 'LD' and 'C' denote LD and Citation labels, while 'F' and 'C' refer to inputs from Facts or Considerations. Models marked with (*) are zero-shot LLMs evaluated on the validation set.

Model	LD-F	LD-C	C-F	C-C	Agg.
Random Baseline	36.2	36.0	24.1	25.7	29.5
Majority Baseline	49.1	49.2	11.8	11.8	19.0
MiniLM	54.7	65.8	9.8	20.8	21.8
DistilmBERT	56.2	65.4	19.6	22.1	30.9
mDeBERTa-v3	55.1	69.8	21.0	17.5	29.1
XLM-R$_{Base}$	57.2	65.9	21.3	23.7	32.8
XLM-R$_{Large}$	56.4	67.9	24.4	**29.1**	**37.1**
X-MOD$_{Base}$	56.6	67.8	20.0	20.6	30.5
SwissBERT$_{(xlm-vocab)}$	56.9	67.3	25.7	23.0	34.8
mT5$_{Small}$	52.2	62.1	13.2	17.9	24.0
mT5$_{Base}$	52.1	61.5	14.0	19.7	25.4
BLOOM$_{560M}$	53.0	61.7	10.7	8.0	15.8
Legal-Swiss-RoBERTa$_{Base}$	57.7	70.5	16.2	20.1	28.0
Legal-Swiss-RoBERTa$_{Large}$	55.9	68.9	25.8	16.3	30.2
Legal-Swiss-LF$_{Base}$	**58.1**	**70.8**	21.4	17.4	29.5
GPT-3.5*	46.6	44.8	25.7	16.7	28.1
LLaMA-2*	45.2	26.6	7.0	8.5	12.5

outperform their smaller counterparts. For example, mT5$_{Base}$ and mT5$_{Small}$ both underperformed DistilmBERT in all configurations. The Legal-Swiss models performed well in LD labels, particularly the Legal-Swiss-LF$_{Base}$, which achieved the highest scores in LD-F and LD-C. However, their weak performance on Citation labels highlights the dataset's complexity, even for domain-specific models. GPT-3.5 and LLaMA-2 underperform fine-tuned models, underlining the need for specialized models for these tasks. The difference is largest in the LD labels where the small fine-tuned models always outperformed LLMs despite their huge parameter number difference (e.g., MiniLM with 118M parameters versus GPT-3.5 with 175B).

Table 2 shows more detailed results on the language specific scores. SwissBERT pretrained with

a focus on German achieved the highest aggregate score in German, but interestingly with scores in Italian being the highest by far in C-F. Models pretrained on CC100 (Conneau et al., 2020) (MiniLM, mDeBERTa, XLM-R and X-MOD) exhibited mixed results in French and Italian, with all models performing best in German. MiniLM, mDeBERTa, and X-MOD showed underperformance in Italian but stronger results in French. In contrast, XLM-R, particularly the large variant, demonstrated robust performance in Italian. mT5 models performed well in French, and the base variant additionally also performed well on Italian. BLOOM was much better in French than in other languages, not surprising given it did not have German and Italian in the pretraining data.

Overall, the connection between the proportion of a language in the pretraining corpus and the model's downstream performance in that language appears weak. While some models showed stronger results in languages with higher representation in their pretraining data, this trend was inconsistent across all models and languages.

6 Conclusions and Future Work

This work introduced a novel Criticality Prediction task to assess the potential influence of SFSC cases on future jurisprudence. Our approach utilizes leading decision citation for a more comprehensive and challenging multilingual evaluation of case law importance compared to existing methods. We also released its multilingual dataset for a community.

We conducted a comprehensive evaluation of our Criticality Prediction task, comparing a range of models, from smaller multilingual models to LLMs

like GPT-3.5. Our findings show that small fine-tuned models consistently outperform zero-shot LLMs, highlighting the importance of task-specific adaptation in legal NLP applications.

Future studies could apply the Criticality Prediction task in other legal contexts by incorporating sources from different jurisdictions and languages. This approach would broaden the research's impact and enhance the model's adaptability across different legal systems.

Limitations

It is very difficult to estimate the importance of a case. By relying on proxies such as whether the case was converted to a leading decision (LD-label) and how often this leading decision was cited (Citation-label), we were able to create labels algorithmically. While we discussed this with lawyers at length and implemented the solution we agreed on finally, this task remains somewhat artificial. Additionally, it is worth noting that LLMs can be sensitive to prompt formats and the order of answer options, leading to inconsistent outputs as highlighted in recent research (Webson and Pavlick, 2022; Elazar et al., 2021). Incorporating varied prompt designs into future studies could strengthen our approach further.

Ethics Statement

While automating case prioritization and identifying leading decisions can greatly benefit legal professionals, there are potential risks associated with deploying such classifiers. One concern is the risk of perpetuating biases present in historical legal trends. For instance, case prioritization decisions should not be influenced by factors such as gender, race, or other protected characteristics. We acknowledge these concerns and will pursue measures to mitigate such biases in future work.

Additionally, there are challenges with reproducibility when using closed models like ChatGPT. Since the internal workings of these models are not fully transparent, results may be difficult to replicate. To promote open science, we have provided comprehensive evaluations of open source multilingual models, aiming to make our findings more accessible and reproducible.

Acknowledgments

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

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

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

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction: A survey of the state of the art. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*,

pages 5461–5469. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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

Maria Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the European Court of Human Rights: looking into the crystall ball. *Proceedings of the Conference on Empirical Legal Studies in Europe 2018*.

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, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

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.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Sturmer, Ilias Chalkidis, Daniel Ho, and Joel Niklaus. 2023. One law, many languages: Benchmarking multilingual legal reasoning for judicial support.

V Sanh. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet

Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint*. ArXiv:2211.05100 [cs].

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We Need to Talk About Random Splits. *arXiv:2005.00636 [cs]*. ArXiv: 2005.00636.

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

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The Multilingual Language Model for Switzerland. *arXiv e-prints*, pages arXiv–2303.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*. ArXiv: 2010.11934.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

A Weighting Formula for Citation-Label

The weighting formula used for the **Citation-Label** is designed to balance the impact of older cases with more recent rulings, ensuring that the ranking reflects both citation frequency and recency. The formula is as follows:

$$score = count \times \frac{year - 2002 + 1}{2023 - 2002 + 1}$$

Where:

- *count* refers to the number of times a particular case is cited in Swiss Federal Supreme Court (SFSC) decisions.

- The year 2002 is the starting year of our dataset.

- The year 2023 is the end point of the our dataset.

We have +1 adjustment in weighting factor $\frac{year - 2002 + 1}{2023 - 2002 + 1}$. This ensures that cases from the year 2002 are still included in the weighted calculation and do not receive a weight of zero.

We did not use more than four bins for the Citation-Label to ensure enough examples also for the smallest class while still adding more granularity compared to the LD-Label.

B Dataset Example

Figure 3 shows an example input, target and corresponding metadata.

C General Dataset Metadata

Table 4 shows metadata for cantons, courts, chambers law-areas and languages.

D Dataset Licensing

The original case data is available from the Swiss Federal Supreme Court[2] and the Entscheidsuche portal[3] was used to download HTML files for each case.

In compliance with the Swiss Federal Supreme Court's licensing policy[4], we are releasing the dataset under a CC-BY-4.0 license. *The link to the dataset will be made available upon acceptance.*

Personally identifying information has already been anonymized by the Swiss Federal Supreme Court in accordance with its anonymization rules[5].

E Zero-shot Prompts

Figures 4 and 5 show the zero-shot prompts we used for the Case Prioritization and Leading Decision Selection tasks respectively.

F Hyperparameters and Package Settings

We used a fixed learning rate of 1e-5 without tuning, running each experiment with three random seeds (1-3) and excluding seeds with high evaluation losses. Gradient accumulation was applied when GPU memory was insufficient to maintain a final batch size of 64. Training employed early stopping with a patience of 5 epochs, based on validation loss. To reduce costs, AMP mixed precision was used where it didn't cause overflows (e.g., mDeBERTa-v3). The max-sequence length was set at 2048 for Facts and 4096 for Considerations. For LLM evaluations, we used the ChatCompletion API for GPT-3.5 (as of June 7, 2023), and ran LLaMA-2 locally with 4-bit quantization.

For the analysis of consideration and fact lengths, we used SpaCy's *en_core_web_sm* for tokenization.

G Resources

The experiments were run on NVIDIA GPUs, including the 24GB RTX3090, 32GB V100, 48GB A6000, and 80GB A100, using approximately 50 GPU days in total.

H Additional Results

Table 5 shows results with standard deviations on the test set. Table 6 shows scores on the validation set.

I Use of AI Assistants

We used ChatGPT and Claude to enhance grammatical correctness and style, and utilized Google Colab's Generate AI feature for some of the dataset analysis.

---

[2] https://www.bger.ch/de/index.htm

[3] https://entscheidsuche.ch/

[4] https://www.bger.ch/files/live/sites/bger/files/pdf/de/urteilsveroeffentlichung_d.pdf

[5] https://www.bger.ch/files/live/sites/bger/files/pdf/Reglemente/Anonymisierungsregeln_2020_def__d.pdf

Input

[Consideraions]: Erwägungen: 1. Angefochten ist der in einem kantonal letztinstanzlichen Scheidungsurteil festgesetzte nacheheliche Unterhalt in einem Fr. 30'000.– übersteigenden Umfang; auf die Beschwerde ist somit einzutreten (Art. 72 Abs. 1, Art. 74 Abs. 1 lit. b, Art. 75 Abs. 1 und Art. 90 BGG). 2. Die Parteien pflegten eine klassische Rollenteilung, bei der die Ehefrau die Kinder grosszog und sich um den Haushalt kümmerte. Infolge der Trennung nahm sie im November 2005 wieder eine Arbeitstätigkeit auf und erzielt mit einem 80%-Pensum Fr. 2'955.– netto pro Monat. Beide kantonalen Instanzen haben ihr jedoch auf der Basis einer Vollzeitstelle ein hypothetisches Einkommen von Fr. 3'690.– angerechnet. Das Obergericht hat zwar festgehalten, der Ehefrau sei eine Ausdehnung der Arbeitstätigkeit kaum möglich, gleichzeitig aber erwogen, es sei nicht ersichtlich, weshalb sie nicht einer Vollzeitbeschäftigung nachgehen könne. Ungeachtet dieses Widerspruches wird das Einkommen von Fr. 3'690.– von der Ehefrau ausdrücklich anerkannt, weshalb den nachfolgenden rechtlichen Ausführungen dieser Betrag zugrunde zu legen ist. [...]

Metadata:
Decision ID: *65aad3f6-33c2-4de2-91c7-436e8143d6ea*
Year: *2007*
Language: *German*
Law Area: *Civil*
LD Label: *Critical*
Citation Label: *Citation-1*
Court: *CH_BGer*
Chamber: *CH_BGer_005,*
Canton: *CH*
Region: *Federation*

Target:
critical-1
Possible LD label: critical, non-critical, Possible citation label: critical-1, critical-2, critical-3, critical-4

Figure 3: Example of an input, target and corresponding metadata

Table 4: Metadata for cantons, courts, chambers, law-areas, and languages

Metadata	Number	Examples
Cantons	1	Federation (CH)
Courts	1	Supreme Court
Chambers	13	CH-BGer-011 CH-BGer-004 CH-BGer-008 CH-BGer-002 CH-BGer-005 CH-BGer-001 CH-BGer-006 CH-BGer-009 CH-BGer-015 ...
Law-Areas	4	Civil, Criminal, Public, Social
Languages	5	German, French, Italian

Criticality Prediction (CP) LD Facts/Consideration

Given the **{facts/considerations}** from the following Swiss Federal Supreme Court Decision:
{INPUT FROM THE VALIDATION SET}
Federal Supreme Court Decisions in Switzerland that are published additionally get the label critical, those Federal Supreme Court Decisions that are not published additionally, get the label non-critical. Therefore, there are two labels to choose from:
- **critical**
- **non-critical**
The relevant label in this case is:

Figure 4: Prompt used for Criticality Prediction (LD Facts/-Consideration). The LLM is tasked with predicting whether a Swiss Federal Supreme Court decision is labeled as critical or non-critical based on the provided facts or considerations

Table 5: Configuration aggregate scores with standard deviations on the test set. The macro-F1 scores are provided.

Model	LD-F	LD-C	C-F	C-C	Agg.
MiniLM	$54.7_{+/-1.9}$	$65.8_{+/-1.6}$	$9.8_{+/-2.8}$	$20.8_{+/-3.0}$	21.8
DistilmBERT	$56.2_{+/-0.5}$	$65.4_{+/-1.7}$	$19.6_{+/-1.1}$	$22.1_{+/-0.4}$	30.9
mDeBERTa-v3	$55.1_{+/-2.0}$	$69.8_{+/-2.8}$	$21.0_{+/-3.6}$	$17.5_{+/-4.4}$	29.1
XLM-R$_{Base}$	$57.2_{+/-1.5}$	$65.9_{+/-3.2}$	$21.3_{+/-1.5}$	$23.7_{+/-1.9}$	32.8
XLM-R$_{Large}$	$56.4_{+/-1.8}$	$67.9_{+/-1.9}$	$24.4_{+/-7.2}$	$29.1_{+/-2.7}$	37.1
X-MOD$_{Base}$	$56.6_{+/-1.8}$	$67.8_{+/-2.9}$	$20.0_{+/-3.0}$	$20.6_{+/-3.5}$	30.5
SwissBERT$_{(xlm-vocab)}$	$56.9_{+/-0.7}$	$67.3_{+/-4.7}$	$25.7_{+/-8.3}$	$23.0_{+/-4.0}$	34.8
mT5$_{Small}$	$52.2_{+/-1.9}$	$62.1_{+/-5.2}$	$13.2_{+/-2.4}$	$17.9_{+/-1.7}$	24.0
mT5$_{Base}$	$52.1_{+/-1.6}$	$61.5_{+/-3.9}$	$14.0_{+/-2.8}$	$19.7_{+/-1.6}$	25.4
BLOOM-560m	$53.0_{+/-1.7}$	$61.7_{+/-4.1}$	$10.7_{+/-3.7}$	$8.0_{+/-3.5}$	15.8
Legal-Swiss-R$_{Base}$	$57.7_{+/-1.6}$	$70.5_{+/-2.3}$	$16.2_{+/-5.8}$	$20.1_{+/-5.6}$	28.0
Legal-Swiss-R$_{Large}$	$55.9_{+/-2.2}$	$68.9_{+/-2.1}$	$25.8_{+/-7.8}$	$16.3_{+/-8.7}$	30.2
Legal-Swiss-LF$_{Base}$	$58.1_{+/-2.1}$	$70.8_{+/-2.9}$	$21.4_{+/-2.9}$	$17.4_{+/-8.6}$	29.5

Criticality Prediction (CP) Citation Facts/Consideration

Given the **{facts/considerations}** from the following Swiss Federal Supreme Court Decision:
{INPUT FROM THE VALIDATION SET}
How likely is it that this Swiss Federal Supreme Court Decision gets cited. Choose between one of the following labels (a bigger number in the label means that the court decision is more likely to be cited):
- **critical-1**
- **critical-2**
- **critical-3**
- **critical-4**
The relevant label in this case is:

Figure 5: Prompt used for Criticality Prediction (Citation Facts/Consideration). The LLM is tasked with predicting how likely a Swiss Federal Supreme Court decision is to be cited, using a four-tiered label (critical-1 to critical-4), based on the provided facts or considerations.

Table 6: Configuration aggregate scores on the validation set. The macro-F1 scores are provided. The highest values are in bold. It is important to note that the scores presented here are calculated as the harmonic mean over multiple seeds.

Model	LD-F	LD-C	C-F	C-C	Agg.
MiniLM	59.1	71.0	14.9	36.9	31.9
DistilmBERT	59.6	70.1	26.3	35.8	41.2
mDeBERTa-v3	60.1	73.0	30.4	36.0	44.0
XLM-R$_{Base}$	60.1	70.5	26.9	38.5	42.6
XLM-R$_{Large}$	60.5	71.7	27.2	39.7	43.3
X-MOD$_{Base}$	57.1	71.0	27.0	33.4	40.6
SwissBERT$_{(xlm-vocab)}$	59.0	72.1	29.4	38.8	44.1
mT5$_{Small}$	54.8	66.1	26.3	32.5	39.2
mT5$_{Base}$	55.7	64.4	24.3	29.3	36.8
BLOOM-560m	52.2	64.3	20.1	21.8	30.7
Legal-Swiss-R$_{Base}$	61.2	**73.6**	27.7	41.0	44.2
Legal-Swiss-R$_{Large}$	**61.8**	73.5	29.8	32.0	42.3
Legal-Swiss-LF$_{Base}$	59.4	72.7	**32.2**	**42.5**	**47.0**