

# ⚖️ *SCALE*: Scaling up the Complexity for Advanced Language Model Evaluation

Vishvaksenan Rasiah<sup>1\*</sup>

Ronja Stern<sup>1\*</sup>

Veton Matoshi<sup>2</sup>

Matthias Stürmer<sup>1,2</sup>

Ilias Chalkidis<sup>3</sup>

Daniel E. Ho<sup>4</sup>

Joel Niklaus<sup>1,2,4\*</sup>

<sup>1</sup>University of Bern <sup>2</sup>Bern University of Applied Sciences

<sup>3</sup>University of Copenhagen <sup>4</sup>Stanford University

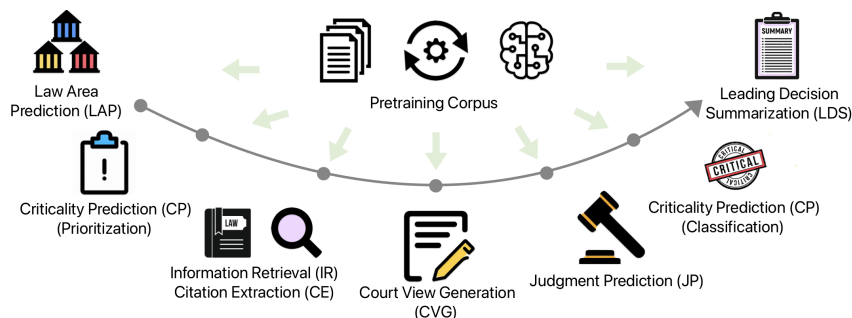


Figure 1: Sequence of tasks for support in the judicial system.

## Abstract

Recent strides in Large Language Models (LLMs) have saturated many NLP benchmarks (even professional domain-specific ones), emphasizing the need for novel, more challenging novel ones to properly assess LLM capabilities. In this paper, we introduce a novel NLP benchmark that poses challenges to current LLMs across four key dimensions: processing *long documents* (up to 50K tokens), utilizing *domain specific knowledge* (embodied in legal texts), *multilingual* understanding (covering five languages), and *multitasking* (comprising legal document to document Information Retrieval, Court View Generation, Leading Decision Summarization, Citation Extraction, and eight challenging Text Classification tasks). Our benchmark comprises diverse legal NLP datasets from the Swiss legal system, allowing for a comprehensive study of the underlying Non-English, inherently multilingual, federal legal system. Despite recent advances, efficiently processing long documents for intense review/analysis tasks remains an open challenge for language models. Also, comprehensive, domain-specific benchmarks requiring high expertise to develop are rare, as are multilingual benchmarks. This scarcity underscores our contribution's value, considering most public models are trained predominantly on English corpora, while other languages remain understudied, particularly for practical domain-specific NLP tasks. Our benchmark allows for testing and advancing the state-of-the-art LLMs. As part of our study, we evaluate several pre-trained multilingual language models on our benchmark to establish strong baselines as a point of reference. Despite the large size of our datasets

\* Equal contribution.

(tens to hundreds of thousands of examples), existing publicly available models struggle with most tasks, even after in-domain pretraining. We publish all resources (benchmark suite, pre-trained models, code) under a fully permissive open CC BY-SA license.

## 1 Introduction

The history of legal Natural Language Processing (NLP) is extensive [3], with remarkable progress recently [47]. Notably, the introduction of datasets containing legal data from various jurisdictions worldwide [75], [14], as well as the development of more domain-specific tasks and benchmarks [38, 57, 86, 10, 44, 71, 93, 20, 31] have significantly contributed to the progress in the field. General benchmarks such as SuperGLUE[97] are saturated and ineffective at differentiating Large Language Models (LLMs). Hence, larger, challenging benchmarks are urgently needed, especially in the domain-specific context. In the context of Switzerland, the availability of only one dataset for evaluating LLMs hampers the assessment of their performance and effectiveness within the country’s diverse linguistic and legal landscape [69, 70]. In this paper, we introduce seven related datasets covering a range of tasks and spanning across five languages within the same overarching jurisdiction. These datasets are derived from 26 cantons and the Swiss Federal Supreme Court (FSCS), each with distinct legal frameworks, in the uniquely multilingual and multi-jurisdictional context of Switzerland. The country’s multiple official languages and a wealth of data for its size, position Switzerland as an exemplary testbed for assessing LLMs in a multilingual and multi-jurisdictional environment. Our assessment concentrates on three classification tasks – Criticality Prediction (CP), Judgment Prediction (JP), and Law Area Prediction (LAP) – an Information Retrieval (IR) task and two generative tasks – Court View Generation (CVG) and Leading Decision Summarization (LDS). To facilitate a comprehensive analysis and provide baselines for future research, we evaluate an array of models on our datasets similar to Hwang et al. [44] or Niklaus et al. [71]. Furthermore, we have pretrained our own Swiss legal models, Legal Swiss RoBERTa<sub>Base/Large</sub> and Legal Swiss Longformer<sub>Base</sub>. Our tasks challenge current models significantly, with the best performing model only achieving an aggregated Macro F1 score of 48.4. ChatGPT was not able to solve the text classification tasks well, considerably lagging behind finetuned models. The results for CP, IR and CVG are particularly underwhelming, seeming rather arbitrary. We invite the research community to develop new methods to tackle these hard tasks. All data employed in this study is in the public domain and is available on the HuggingFace Hub under a CC BY-SA license<sup>2</sup>.

This paper makes three contributions. First, we present seven public multilingual datasets containing Swiss legal documents. Second, we release two large, in-domain pretraining datasets, and pretrain three new models - Legal-Swiss-RoBERTa<sub>Base/Large</sub> and Legal-Swiss-LongFormer<sub>base</sub>. Third, we evaluate multilingual baselines on our datasets and compare them to our models. Although in-domain pretraining improves performance, significant room for improvement remains in most tasks.

## 2 Related Work

We briefly discuss prior work on benchmarks for long documents, domain specificity, multilinguality, and multitasking. Additional task-specific related work is presented in Appendix F.

**Long Documents** SCROLLS consists of summarization, Question Answering (QA), and Natural Language Inference (NLI) tasks with example inputs typically in the thousands of English words [88]. MuLD is a set of six tasks (twice QA, style change detection, classification, summarization, and translation) where each input is at least 10K tokens, with some up to almost 500K tokens [43].

**Domain Specificity** The BLUE benchmark [76] contains five tasks over ten datasets for biomedical and clinical texts. CBLUE [111] is a Chinese biomedical benchmark with eight understanding tasks including Named Entity Recognition (NER), information extraction, diagnosis normalization, Text Classification (TC), QA, intent classification, and semantic similarity. LEXGLUE covers six predictive tasks over five datasets made of documents in English from the US, EU, and Council of Europe [19]. LEXTREME is a multi-lingual and multi-task benchmark for the legal domain [71].

---

<sup>2</sup><https://huggingface.co/reds>

LegalBench [31] covers zero-shot and few-shot Language Model (LM) evaluation for diverse realistic legal tasks in English. LBOX OPEN [44] consists of five legal tasks from South Korea.

**Multilinguality** XTREME [39], designed to evaluate cross-lingual generalization, includes six tasks across ten datasets, covering 40 languages. Some datasets were cross-lingual, others were extended via professional and automatic translations. XTREME-UP expands XTREME, emphasizing the evaluation of multilingual models in a few-shot setting for user-centric tasks [83]. It covers 88 under-represented languages such as Swahili, Burmese, or Telugu where only few datasets exist.

**Multitasking** GLUE [96], an early benchmark of sentence NLU tasks evaluating general-purpose neural LMs, quickly became obsolete due to advanced models like BERT [25]. Its updated version, SUPERGLUE [97], introduced new tasks challenging for machines yet solvable by humans. MMLU features only zero-shot and few-shot learning tasks [37], containing about 16K multiple-choice questions divided into 57 subtasks, spanning subjects in the humanities, social and hard sciences, etc. CLUE [106] is the first Chinese language multitask benchmark that includes single sentence classification, sentence pair classification, and machine reading comprehension. BIG-Bench [91] consists of 204 language tasks created by 450 authors from 132 institutions. The tasks cover topics such as linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development. HELM [59] is a multi-metric benchmark covering seven metrics and seven targeted evaluations and involves 42 test scenarios with a large-scale evaluation of 30 LMs.

### 3 Background on the Swiss Legal System

Switzerland consists of 26 cantons, each with its own jurisdiction and different organisation of the courts. The Swiss Federal Supreme Court (FSCS) is the highest instance of Swiss jurisprudence and the final instance for cases at the federal criminal court, the federal administrative court, the federal patent court and any cantonal court. Its decisions contribute to the development of the law and to its adaptation to changing circumstances. The FSCS is divided into seven different divisions, mainly distinguishing between public, penal and civil law [11]. While all cases before the Supreme Court are Federal Supreme Court Decisions (FSCD), only a few are additionally marked as Leading Decision (BGE) and are published separately. Cantonal court proceedings start at the lowest instance and can be appealed higher. The stages of appeal depend on the canton and the area of law.

### 4 SCALE: The Datasets

Table 1 introduces our eleven datasets. Data was collected from 26 cantons (in addition to federal decisions), 184 courts, 456 chambers, four main law areas, and five languages as seen in Appendix H Table 8. There are significant differences in the availability of documents across cantons and courts.<sup>3</sup> Most courts are monolingual, but there are cantons where multiple languages are used in documents. In addition to the decision-based datasets, we also provide a collection of approx. 35K laws from cantonal and federal jurisdictions in Switzerland. While most FSCD are written in German, French is more common for cantonal cases. We split all downstream datasets into train (until 2015), validation (2016-2017) and test (2018-2022) split based on the dates. We opted for a relatively large test split, because LLMs seem to need relatively little training data [9]. A large test set allows future researchers to perform detailed studies on the performance over multiple years (including the three special years 2020 to 2022 of the COVID-19 pandemic). This large date

Table 1: Overview over all datasets and their multilingualism: Abbreviations: **C**antonal, **F**ederal, **F**acts, **C**onsiderations Column Fac and Cons report the mean token lengths. Sections Facts and Considerations are not available for Ruling Summarization, Legislation and Rulings due to different format, thus mean token length for the full text is reported and marked with \*.

Name	Level	Total	DE	FR	IT	RM	EN	Fac	Cons
Rulings	Cant + Fed	638K	320K	247K	71K	-	180	-	*7K
Leading Decisions	Fed	21K	14K	6K	1K	-	-	689	3K
Legislation	Cant + Fed	36K	18K	11K	6K	534	207	-	*7K
Doc2Doc IR	Fed	141K	87K	46K	8K	-	-	847	3K
Citation Extraction	Fed	131K	85K	38K	8K	-	-	-	204
Criticality	Fed	139K	85K	45K	8K	-	-	828	3K
Law Area	Cant + Fed	329K	127K	156K	46K	-	-	2K	4K
Judgment Prediction	Cant + Fed	329K	160K	128K	41K	-	-	2K	4K
Court View	Cant + Fed	404K	197K	163K	44K	-	-	2K	5K
Court View Origin	Fed	270	49	221	-	-	-	1K	6K
Leading Decision Summarization	Fed	18K	12K	5K	835	-	-	-	*3K

<sup>3</sup>The FSCS is the only court where we have complete data, since all decisions since 2007 has been published.

difference between the youngest training examples (from 2015) and the youngest test examples (2022) may be another reason for models struggling on these tasks (see section 6). The source data is highly curated by established Swiss institutions such as courts and administrative bodies (the courts spend approx. 45 per case for manual anonymization) [69].

#### 4.1 Database Creation Pipeline

Every day, new cases are published on Entscheidsuche.ch, enabling us to fetch new documents daily (see Figure 2). (1) We scrape all files found on Entscheidsuche.ch, which include metadata of every court’s folder. Only case documents that do not already exist locally in our database are sent through the pipeline. (2) BeautifulSoup / tika-python library is used to extract text from HTMLs / PDFs. (3) We extract the corresponding language with the fastText language identification tool [28]. Each case must have an appropriate language for further extraction tasks. (4) A cleaner removes any strange patterns or redundant text to avoid errors for further extraction tasks. (5) Cases are split into the sections header, facts, considerations, rulings, and footer using a set of regex patterns. (6) To extract judgment outcome, a set or combination of words is defined for each predefined outcome. Since those indicators are not exclusive to this context, it is crucial to consider only the ruling section of a case to avoid false positives. Therefore, successful judgment outcome extraction is dependent on precise section splitting. (7) BGE and law citations are procured either via Regex (cantonal) or BeautifulSoup (federal). The FSCS labels all citations with an HTML tag, ensuring a high quality of citations for federal cases.

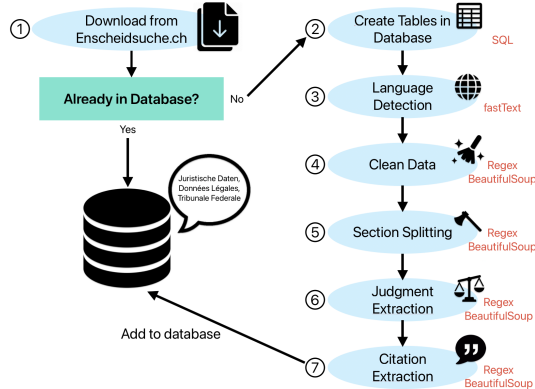


Figure 2: Database Creation Pipeline

#### 4.2 Pretraining

**Legislation** The Swiss Legislation dataset comprises 35.7K legislative texts (182M tokens) distributed across five languages: German, French, Italian, Romansh, and English (see Figure 6). Figure 7 details its coverage of federal, cantonal, and inter-cantonal legislation on a broad array of legal topics including public health, education, civil rights, societal matters, energy, environment, infrastructure, and visa regulations. It also includes instances of the same legislation texts across different languages, useful for enhancing the multilingual capabilities of legal LMs.

**Rulings** The Swiss Rulings dataset is a comprehensive collection of Swiss court rulings designed for pretraining purposes. It consists of 638K cases (3.3B tokens) distributed across three languages: German (319K), French (247K) and Italian (71K). Spanning several decades and covering multiple areas of law, this dataset provides an extensive representation of Swiss law practice.

#### 4.3 Text Classification

We work with eight different configurations, built from the LAP, CP and JP datasets. While the Law Area and Judgment Prediction datasets include both federal and cantonal cases, the Criticality dataset considers only FSCD. All tasks presented in this section involve Single Label Text Classification (SLTC), which required either extracting or defining labels. For each of the tasks we both consider the facts and the considerations as input. The facts represent the most similar available proxy to the complaints, useful for predictive tasks. The considerations as input make the tasks considerably easier, since they include the legal reasoning. These tasks can be used as post-hoc analyses for verification (e.g., in judgment prediction whether the made judgment is congruent with the given reasoning).

**Law Area Prediction** The **Law Area** label was established by associating a law-area to each chamber where a case was adjudicated. Using metadata from Entscheidsuche.ch, a lawyer helped define the law areas for each chamber, resulting in chambers being classified into one of four main law-area categories (civil, public, criminal and social law) and 12 sub-law-areas. Due to many

Table 2: Task Configurations. Label names are *Critical* (C), *Non-critical* (NC), *Critical-1* (C1) to *Critical-4* (C4), *Approval* (A), *Dismissal* (D). For Law-Sub-Area we reported only the two most common labels *Substantive Criminal* (SC), *Criminal Procedure* (CP), and the two least common *Intellectual Property* (IP), *Other Fiscal* (OF). Abbreviations: **Validation**, **Considerations**, **Facts**

Task Name	Train	Labels Train				Val	Labels Val				Test	Labels Test			
BGE-Fac	<b>75K</b>	C	NC			C	NC			<b>26K</b>	C	NC			
BGE-Con	<b>91K</b>	3K	72K	-	-	12K	580	13K	-	-	948	25K	-	-	
Citation-Fac	<b>2.5K</b>	C-1	C-2	C-3	C-4	<b>563</b>	C-1	C-2	C-3	C-4	<b>725</b>	C-1	C-2	C-3	C-4
Citation-Con	<b>2.5K</b>	782	626	585	513	<b>563</b>	186	152	131	94	<b>725</b>	137	177	224	187
Judgment-Fac	<b>197K</b>	D	A			<b>37K</b>	D	A			<b>94K</b>	D	A		
Judgment-Con	<b>188K</b>	135K	62K	-	-	<b>37K</b>	27K	11K	-	-	<b>94K</b>	67K	27K	-	-
Law-Sub-Area-Fac	<b>10K</b>	SC	CP	IP	OF	<b>9K</b>	SC	CP	IP	OF	<b>3K</b>	SC	CP	IP	OF
Law-Sub-Area-Con	<b>8K</b>	3K	3K	6	2	<b>9K</b>	2K	1K	11	1	<b>3K</b>	1K	509	5	1
		2K	1K	6	2	<b>7K</b>	2K	750	11	1	<b>3K</b>	885	401	3	1

chambers operating in various law areas, it was not always feasible to assign a single law area label to each chamber. Particularly for the more detailed sub-law areas, where several chambers could not be uniquely linked, this resulted in a small subset of cases with the subset label. Initial results on the full dataset including the four main law areas showed that current models achieve near perfect accuracy, which is why we only consider the smaller filtered dataset of sub areas for this benchmark.

**Judgment Prediction** We created the **Judgment** label by extracting the judgment outcome with regex patterns and assign a binary label with two classes: approval and dismissal, similar to [69]. For partially approved or dismissed judgments, we labeled them as approval or dismissal, respectively.

**Criticality Prediction** We quantified **Criticality** in two ways: First, the **BGE-Label** is binary: *critical* and *non-critical*. FSCD are labeled as *critical* if additionally published as Leading Decision (see Section 3). To achieve this, we extracted the FSCD file names from the headers of BGE cases using regex patterns. Cases not found in the header of a BGE were labeled as *non-critical*.<sup>4</sup> Second, to create a more precise adaptation of the BGE-Label, we developed the **Citation-Label**, which involved counting all citations of BGE in all FSCD cases. The BGE frequency was weighted based on recency, with older citations receiving a smaller weight:  $score = count * \frac{year-2002+1}{2023-2002+1}$ . This resulted in a ranking of BGEs, which were then divided into four categories of criticality *critical-1* to *critical-4*. We used the 25, 50 and 75% quartiles as separation for our four classes.

#### 4.4 Text Generation

**Court View Generation** Clerks and judges dedicate a significant portion of their time to preparing considerations for court cases - approximately 50% in penal law and as much as an estimated 85% in other areas of law [69]. Crafting considerations is arguably the central task of a judge’s role, requiring intricate legal knowledge of applicable legislation, caselaw and legal analysis and advanced reasoning skills to connect this myriad of information. The complexity of this task, especially in the Supreme Court, is reflected in the average appointment age of judges being 50 years<sup>5</sup>, underscoring the length and difficulty of their professional journey. Given these time and expertise demands, the necessity of the court view generation task emerges, aiming to create case considerations from the facts. Generating court views is challenging due to several reasons: Both the facts (input) and the considerations (output) can be long and complex. Current models, constrained by their limitations in handling long context, often fail to fully process this extensive input. This shortcoming, when coupled with the input’s inherent complexity, underscores the deficiencies of current models. Aiming to overcome these limitations, we present a novel Court View Generation dataset containing over 400K cases, covering a diverse range of legal scenarios. With an average length of 1522 tokens for the facts and 4673 tokens for the considerations, this dataset provides a challenging benchmark for models to generate coherent and accurate case considerations from legal facts. Furthermore, we provide a Court View origin dataset featuring federal rulings, enriched with data from the lower courts, including their facts and considerations, as well as those of the federal court. This provides a multilevel judicial

<sup>4</sup> Since we have all BGE but not all FSCD, there are missing *critical* cases.

<sup>5</sup> Mit 28 Jahren Mitglied des Bundesgerichts, SonntagsZeitung, December 19, 2019, p. 24

perspective, contributing to a more comprehensive understanding of case progression and further augmenting the challenge of court view generation.

**Leading Decision Summarization** BGE are crucial in the Swiss legal system, often cited to clarify legislative gaps. Access to their summaries simplifies searching and understanding key concepts, the most important citations, and main themes. In the LDS dataset, we include 18K BGE with their summaries, penned by FSCS clerks and judges.

#### 4.5 Information Retrieval

In our IR task, we structured our data into queries, qrels and corpus (see Appendix 21). We use all Swiss legislation and leading decisions as the corpus, while cases from the FSCS in German, French, and Italian serve as queries. The mean token length of our queries is significantly longer than that in other IR benchmarks due to our use of entire documents as queries (see Appendix H Table 1). The goal is to find laws and decisions cited within a given case. We take the facts as a proxy for an appeal being written by the lawyer. Our ground truth is based on citations found in the considerations. We find relevant laws and decisions by extracting cited law articles and decisions from the Swiss legislation and leading decisions datasets, respectively. Document lengths mirror tasks like EU2UK [16]. As laws are written in all three official languages, we end up with cross-lingual query-corpus pairs. Those pairs are then logged as qrels. Long document lengths and cross-lingual aspect may pose challenges for retrieval models. In total we have 10K documents, 101K queries and 2K qrels which results in an average of 19 relevant documents per query.

#### 4.6 Citation Extraction

The FSCS annotates citations with special HTML tags, which we employed to create a Token Classification dataset for Citation Extraction (CE). CE is very complicated to encode with regexes due to the extensive citation rules in legal text. However, with a transformer-based model it can be solved very well (MiniLM achieved over 95 macro F1). For brevity we do not present experiments on this dataset but release the dataset and the trained model as a resource to the community.<sup>6</sup>

#### 4.7 The Big Picture

The pretraining corpus and our seven datasets JP, LAP, CP, IR, CVG, LDS, and CE form a unified framework that resembles an artificial judicial system (see Figure 1). Pre-training serves as the foundation, equipping models with the ability to specialize in the respective tasks and thereby enhancing their performance. The remaining tasks, all interconnected, focus on the output of the judicial system. Superior performance in one task can bolster the effectiveness of the others. LAP facilitates routing decisions to the correct chambers inside a court. CP enables courts allocating resources and setting priorities. IR identifies the relevant documents for a case, facilitating the JP and CVG tasks, which predicts the case’s outcome and synthesizes a coherent text to explain the decision’s rationale. CE automatically extracts citations to enrich the final decision before publication. LDS condenses the reasoning into a short summary. Together, these tasks model (albeit still primitively) the flow of the judicial system end-to-end, the first of this kind, to the best of our knowledge.

## 5 Experiments

In this section, we present the pretraining of our legal models and describe the experimental setup for each of the tasks. Besides BLOOM [85] and mT5 [107], there is a scarcity of available multilingual LLMs, with most recent work pretraining on English only. We excluded BLOOM from our evaluation due to its lack of pretraining on German text. Additionally, the extensive context width of our data posed practical challenges, rendering it infeasible to experiment with models larger than 2B parameters. Consequently, our evaluation was limited to models up to the size of mT5<sub>Large</sub> (1.2B).

---

<sup>6</sup><https://huggingface.co/organizations/rcds>

Table 3: Models: BS is short for batch size. InLen is the maximum input length the model has seen during pretraining. # Parameters is the total parameter count (including the embedding layer). Our models were built upon the pre-trained RoBERTa/Longformer. SwissBERT was further trained from pre-trained X-MOD. Utilizing three language adapters with X-MOD and SwissBERT led to fewer parameters and languages.

Model	Source	InLen	# Parameters	Vocab	# Steps	BS	Corpus	# Langs
MiniLM	Wang et al. [99]	512	118M	250K	1M	256	2.5TB CC100	100
DistilBERT	Sanh et al. [84]	512	135M	120K	n/a	< 4000	Wikipedia	104
mDeBERTa-v3	He et al. [34, 33]	512	278M	128K	500K	8192	2.5TB CC100	100
XLM-R <sub>Base/Large</sub>	Conneau et al. [24]	512	278M/560M	250K	1.5M	8192	2.5TB CC100	100
X-MOD <sub>Base</sub>	Pfeiffer et al. [79]	512	299M	250K	1M	2048	2.5TB CC100	3 (81)
SwissBERT (XLM vocab)	Vamvas et al. [95]	512	299M	250K	364K	768	Swissdix	3 (4)
mT5 <sub>Small/Base/Large</sub>	Xue et al. [107]	1024	300M/580M/1.2B	250K	1M	1024	mC4 (CC)	101
Legal-Swiss-R <sub>Base</sub>	ours	512	184M	128K	1M	512	CH Rulings/Legislation	3
Legal-Swiss-R <sub>Large</sub>	ours	512	435M	128K	500K	512	CH Rulings/Legislation	3
Legal-Swiss-LF <sub>Base</sub>	ours	4096	208M	128K	50K	512	CH Rulings/Legislation	3

## 5.1 Pretraining Legal Models

As part of this study, we release two multi-lingual legal-oriented PLMs, dubbed Legal-Swiss-RoBERTa and a Longformer, dubbed Legal-Swiss-LF<sub>Base</sub> trained on Swiss rulings and legislation additional to EUR-LEX data [72].<sup>7</sup> For the newly released Legal-Swiss-RoBERTa models we followed a series of best-practices in LM development literature described in more detail in Appendix G. We make all our models publicly available alongside all intermediate checkpoints (every 50K/10K training steps for RoBERTa/Longformer models) on the Hugging Face Hub.<sup>8</sup> Due to limited resources, we were not able to pretrain a large generative model and leave this to future work.

## 5.2 Text Classification

For our TC tasks, namely LAP, JP, and CP, we adopted the LEXTREME benchmark setup [71]. To be fair across languages, we computed the harmonic mean over all languages to arrive at the score per task. The final aggregate score is the harmonic mean over all task scores [71]. Furthermore, we evaluated SwissBERT [95], selected for its Swiss-specific pretraining data; X-MOD [79] and our custom pre-trained models, namely Legal Swiss RoBERTa and Legal Swiss Longformer (see Table 3). Finally, we evaluated ChatGPT (gpt3.5-turbo as of June 7 2023) by following the experimental setup as described in [12]. We utilized the ChatCompletion API to provide one instruction and example at a time as input. In contrast to Chalkidis [12], we randomly selected 1000 samples (if available) from the validation set instead of the test set<sup>9</sup> (to not leak the test set for future evaluations) and kept only those samples that did not exceed the model’s maximum limit of 4096 tokens. Our experiments were focused solely on zero-shot classification due to the long input lengths.

## 5.3 Text Generation

For our experiments in text generation, we employed mT5 [107], a multilingual encoder-decoder model available in small, base, and large variants. We adopted an effective batch size of 16 using gradient accumulation when necessary. For evaluation we reported BERTScore [112], BLEU [74], METEOR [5], and ROUGE [60]. Each individual metric possesses inherent weaknesses [112], hence the necessity for employing multiple metrics for a more comprehensive assessment.

**Court View Generation** We faced limitations due to lengthy input (avg. 1522 tokens) and output (avg. 4673 tokens) for CVG. To tackle these, we truncated input facts to 2048 tokens and output considerations to 512 tokens. In 90% of the cases, we were able to retain the complete facts. This decision was due to the substantial resources required for longer sequences and the task’s inherent difficulty, expected to be challenging even with only 512 output tokens. Furthermore, due to the large amount of test data and compute constraints, we limited our evaluation to a 1K-instance subset. For the origin dataset, the input was split evenly between origin facts and origin considerations.

**Leading Decision Summarization** For our LDS experiments, similar to CVG, we dealt with substantial input text (avg. 3081 tokens), but shorter output text (avg. 168 tokens). To address these

<sup>7</sup>[https://huggingface.co/datasets/joelito/eurlex\\_resources](https://huggingface.co/datasets/joelito/eurlex_resources)

<sup>8</sup><https://huggingface.co/joelito>

<sup>9</sup>The limitation to 1000 examples primarily served to keep the costs limited. We spent approx. 40\$.

computational challenges, we truncated the input text to a maximum of 4096 tokens and the output texts to a length of 256 tokens, allowing us to preserve full output in over 80% of the cases.

## 5.4 Information Retrieval

Finding relevant legal references for FSCD is a demanding task (a) due to complexities of legal language, (b) multilinguality and (c) long nature of the documents. We investigate the task of multilingual Doc2Doc IR in the legal domain using our new Doc2Doc IR dataset. This dataset includes FSCD, each marked with a unique identifier for law citations and BGE. Since the complexity of the task increases with the number of documents, we expect performance to decrease with additional data. We conducted an ablation study by making minor adjustments to the datasets to understand their impact on model performance (see Appendix I). We used BM25, which scales well to long documents but struggles with contextual processing and handling different languages[82]. Neural approaches like Sentence-Bert (SBERT) [81] show promising results, but perform worse on long input texts due to loss of context after truncation. In addition to standard training (using only positive examples), we explore a training that involves using hard negatives [109]. Cross Encoder models [98] were excluded, due to their high computational cost. Additional to Normalized Discounted Cumulative Gain (NDCG) [100] we use the Capped Recall@k score [93].<sup>10</sup>

## 6 Results

### 6.1 Text Classification

Table 4: Results on the Text Classification datasets. Macro F1 score is reported. The 'F' or 'C' following the dash represents input based on 'Facts' or 'Considerations' respectively. 'CPB' and 'CPC' refer to the CP task using BGE and Citation labels, respectively, while 'SLAP' denotes Sub Law Area Prediction. Note: Seeds that yielded very high evaluation losses were considered as failed and therefore excluded from the analysis.

Model	CPB-F	CPB-C	CPC-F	CPC-C	SLAP-F	SLAP-C	JP-F	JP-C	Agg.
MiniLM	54.7	65.8	9.8	20.8	59.7	61.1	58.1	78.5	32.4
DistilBERT	56.2	65.4	19.6	22.1	63.7	65.9	59.9	75.5	42.1
mDeBERTa-v3	55.1	<b>69.8</b>	21.0	17.5	63.8	59.3	60.6	77.9	40.2
XLM-R <sub>Base</sub>	57.2	65.9	21.3	23.7	<b>67.2</b>	73.4	60.9	79.7	44.6
XLM-R <sub>Large</sub>	56.4	67.9	24.4	<b>28.5</b>	65.1	<b>78.9</b>	60.8	<b>80.9</b>	<b>48.4</b>
X-MOD <sub>Base</sub>	<b>56.6</b>	67.8	20.0	20.6	63.9	64.4	60.5	79.1	41.9
SwissBERT (xlm-vocab)	56.2	67.3	<b>25.7</b>	23.0	62.2	75.1	<b>61.4</b>	79.4	44.6
Legal-CH-R <sub>Base</sub>	57.6	<b>72.8</b>	<b>23.1</b>	22.5	<b>81.6</b>	83.0	64.0	86.4	<b>46.9</b>
Legal-CH-R <sub>Large</sub>	57.4	70.8	21.3	<b>23.3</b>	80.4	<b>84.9</b>	62.8	<b>87.0</b>	46.2
Legal-CH-LF <sub>Base</sub>	<b>58.1</b>	70.8	21.4	17.4	80.2	83.5	<b>65.4</b>	86.4	42.8

We show results in Table 4 and more details with standard deviations in Table 12. As expected, models with more parameters generally perform better, with XLM-R<sub>Large</sub> emerging on top. Our pre-trained model, Legal-ch R<sub>Base</sub>, outperformed XLM-R<sub>Base</sub>, suggesting that domain-specific pre-training can lead to significant improvements in model performance. Overall, our pre-trained models showed better aggregated results compared to other models. However,

contrary to expectations, our Legal Swiss RoBERTa<sub>Large</sub> model performed worse than its base model XLM<sub>Large</sub>. Because of the high weight to outliers allotted by the harmonic mean, Legal-ch-R<sub>Large</sub> is penalized a lot by its relatively low performance on CPC-C in comparison to XLM-R<sub>Large</sub>. This result possibly suggests that extensive pre-training is more important than model size, a result consistent with the findings of Liu et al. [61] and Touvron et al. [94]. Despite extra training on extensive data,

Table 5: Results of Court View Generation task. 'In Len' denotes input length in tokens. **Bold**: best within model; underlined: best overall.

Model	In Len ↑	BERT	BLEU	MET	R1 / R2 / RL
mT5 <sub>Large</sub>	2048	<b>75.74</b>	<b>66.92</b>	<b>34.44</b>	<b>34.91 / 15.58 / 33.53</b>
mT5 <sub>Large</sub>	1024	75.56	66.68	34.02	34.26 / 14.72 / 32.87
mT5 <sub>Large</sub>	512	75.27	66.12	33.48	33.61 / 14.26 / 32.21
mT5 <sub>Base</sub>	2048	75.01	65.48	32.89	33.23 / 13.57 / 31.89
mT5 <sub>Base</sub>	1024	<b>75.15</b>	<b>65.73</b>	<b>33.15</b>	<b>33.49 / 13.96 / 32.18</b>
mT5 <sub>Base</sub>	512	74.89	65.55	32.66	32.66 / 13.16 / 31.35
mT5 <sub>Small</sub>	2048	<b>74.13</b>	<b>63.97</b>	<b>30.96</b>	<b>31.29 / 11.01 / 29.90</b>
mT5 <sub>Small</sub>	1024	74.00	63.70	30.68	31.05 / 10.77 / 29.64
mT5 <sub>Small</sub>	512	73.92	63.83	30.57	30.58 / 10.35 / 29.20

<sup>10</sup>The Capped Recall@k is computed as the proportion of relevant documents for a specific query, retrieved from the top k scored list of documents generated by the model. This is a good representation of model success in our specific task, as each query has multiple relevant documents without a need for intra-document ranking.



Legal-ch-LF did not surpass the hierarchical Legal-ch-R-Base model. ChatGPT clearly lags behind finetuned models, underlining the necessity of specialized models for these tasks (see Table 13 for details). The difference is largest in the JP and Sub Law Area Prediction (SLAP) tasks where the finetuned models are best.

## 6.2 Text Generation

**Court View Generation** We observed a clear trend in the performance of models across all scores, with larger models consistently outperforming smaller ones (see Table 5). However, the increase in performance with longer input was found to be minimal, and in some cases, counterproductive. Generally, the generated text displayed stylistic authenticity, resembling typical legal language. However, it often lacked logical coherence in terms of its content, which underscores the current models’ limited capacity to fully grasp the complexities of generating coherent court views. In numerous court cases, we observed a pattern where the target considerations contained similar paragraphs. These paragraphs were often predicted with a relatively good ability (see examples in Table 14). The results derived from the origin dataset were less conclusive (see Table 11). Likely due to the smaller dataset size, the outputs were found to be less indicative of the model’s capabilities.

### Leading Decision Summarization

As expected, results in Table 6 for LDS task performance revealed two trends. Firstly, an increase in input length led to better scores across models. Secondly, larger models tend to outperform smaller ones, although the differences between the base and the large model are not sharply outlined. The quality of the generated text demonstrated a good stylistic imitation of legal language and a more consistent logical coherence compared

Table 6: Results of Leading Decision Summarization task. ‘In Len’ denotes input length in tokens. **Bold**: best within model; underlined: best overall.

Model	In Len ↑	BERT ↑	BLEU ↑	MET ↑	R1 / R2 / RL ↑
mT5 <sub>Large</sub>	4096	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>
mT5 <sub>Large</sub>	2048	<b>73.10</b>	<b>27.21</b>	<b>21.88</b>	<b>31.47 / 12.22 / 29.94</b>
mT5 <sub>Large</sub>	512	70.67	26.89	18.31	24.76 / 6.15 / 23.48
mT5 <sub>Base</sub>	4096	<b>73.33</b>	<b>30.81</b>	<b>23.50</b>	<b>32.43 / 12.78 / 30.87</b>
mT5 <sub>Base</sub>	2048	72.45	30.13	21.94	30.09 / 10.79 / 28.71
mT5 <sub>Base</sub>	512	70.60	27.10	18.31	24.72 / 6.15 / 23.55
mT5 <sub>Small</sub>	4096	<b>72.04</b>	<b>28.68</b>	<b>21.29</b>	<b>29.61 / 10.31 / 28.12</b>
mT5 <sub>Small</sub>	2048	71.38	24.64	19.28	27.88 / 9.19 / 26.54
mT5 <sub>Small</sub>	512	69.66	20.73	15.95	22.91 / 5.36 / 21.85

to the CVG task (see examples in Table 15).

## 6.3 Information Retrieval

Table 7 presents our evaluations, revealing a consistent inability to retrieve the majority of relevant documents, even when k is set to 100. Lexical models generally outperformed others, despite not optimizing hyperparameters for BM25 [17]. Surprisingly, despite the prevalence of German in our dataset, a French language analyzer (used for stemming and stop-word removal) demonstrated superior performance. For SBERT truncation led to context loss, negatively affecting scores – an issue not encountered with lexical models. Training SBERT models using Multiple Negative Ranking Loss [35] yielded significant performance improvement, with the use of hard negative examples proving advantageous. SBERT evaluation on single languages, denoted as DE, FR, and IT, revealed its inability to perform consistently across all languages, which could be caused by the training set consisting of more German than French or Italian documents. More experiments are detailed in Appendix H Table 9 and 10. Overall, our study exposes limitations of models in dealing with multilingualism, long documents, and legal texts, areas relatively underexplored in previous research. These findings provide a basis for the IR community to innovate strategies for these challenges.

Table 7: Results on Information Retrieval with best scores per section in **bold**. Abbreviations: **distil**useBase-multilingual-cased-v1, joelito/swiss-legal-roberta<sub>base</sub>

Model	RCap@ 1 / 10 / 100 ↑	NDCG@ 1 / 10 / 100 ↑
BM25 (fr lang analyzer)	<b>11.37 / 7.74 / 16.54</b>	<b>11.37 / 8.34 / 11.51</b>
SBERT distil	0.90 / 0.75 / 2.64	2.06 / 1.70 / 3.31
SBERT distil + pos	<b>4.40 / 3.92 / 12.64</b>	<b>10.11 / 8.76 / 16.16</b>
SBERT distil + pos + h-neg	3.97 / <b>4.46 / 13.36</b>	9.12 / <b>9.21 / 16.87</b>
SBERT swiss + pos	3.97 / 3.47 / 12.28	9.12 / 7.76 / 15.16
SBERT distil eval on de queries	<b>4.22 / 4.49 / 15.21</b>	<b>8.21 / 8.15 / 15.86</b>
SBERT distil eval on fr queries	1.88 / 2.20 / 9.19	5.77 / 6.22 / 13.94
SBERT distil eval on it queries	0.22 / 0.24 / 0.79	5.43 / 5.74 / 11.44

## 7 Conclusions

We present SCALE, an end-to-end benchmark of seven datasets for the Swiss legal system and evaluate multilingual models as a reference point, and showing low performance of ChatGPT. Even our in-domain pretrained models struggle greatly in most of our tasks, especially the most challenging (CVG and IR), posing opportunities for researchers to improve models and measure progress.

## Acknowledgements

We greatly appreciate Google’s generous support of TPUs v3-8 machines for pretraining the models. This work has been supported by the Swiss National Research Programme “Digital Transformation” (NRP-77)18 grant number 187477. This work is also partly funded by the Innovation Fund Denmark (IFD)19 under File No. 0175-00011A. We would like to thank Daniel Kettiger, Magda Chodup, and Thomas Lüthi for their legal advice, Mara Häusler for hints regarding the criticality label, Adrian Jörg and Marco Buchholz for help in coding the data extraction pipeline, and Entscheidsuche.ch and the Federal Supreme Court of Switzerland for providing data and advice.

## References

- [1] Ehsan Aghaei, Xi Niu, Waseem Shadid, and Ehab Al-Shaer. SecureBERT: A Domain-Specific Language Model for Cybersecurity. In Fengjun Li, Kaitai Liang, Zhiqiang Lin, and Sokratis K. Katsikas, editors, *Security and Privacy in Communication Networks*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 39–56. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-25538-0\_3.
- [2] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel PreoŃiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93, October 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.93. URL <https://peerj.com/articles/cs-93>. Publisher: PeerJ Inc.
- [3] Kevin D. Ashley. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017. doi: 10.1017/9781316761380.
- [4] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018.
- [5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL <https://aclanthology.org/W05-0909>.
- [6] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv:1903.10676 [cs]*, September 2019. URL <http://arxiv.org/abs/1903.10676>. arXiv: 1903.10676.
- [7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*, December 2020. URL <http://arxiv.org/abs/2004.05150>. arXiv: 2004.05150.
- [8] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL <https://aclanthology.org/2020.acl-main.463>.
- [9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prfulla 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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [10] Tobias Brugger, Matthias Stürmer, and Joel Niklaus. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset, May 2023. URL <http://arxiv.org/abs/2305.01211>. arXiv:2305.01211 [cs].
- [11] Schweizerisches Bundesgericht. The paths to the swiss federal supreme court. [https://www.bger.ch/files/live/sites/bger/files/pdf/en/BG\\_Brosch%C3%BCreA5\\_E\\_0n1.pdf](https://www.bger.ch/files/live/sites/bger/files/pdf/en/BG_Brosch%C3%BCreA5_E_0n1.pdf), 2019. Accessed: 2023-04-27.
- [12] Ilias Chalkidis. ChatGPT may Pass the Bar Exam soon, but has a Long Way to Go for the LexGLUE benchmark. *ArXiv*, abs/2304.1, 2023.
- [13] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1424. URL <https://www.aclweb.org/anthology/P19-1424>.
- [14] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation. In *Proceedings of the Natural Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2209. URL <https://www.aclweb.org/anthology/W19-2209>.
- [15] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]*, October 2020. URL <http://arxiv.org/abs/2010.02559>. arXiv: 2010.02559.
- [16] Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalous, and Prodromos Malakasiotis. Regulatory compliance through doc2doc information retrieval: A case study in eu/uk legislation where text similarity has limitations, 2021.
- [17] Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalous, and Prodromos Malakasiotis. Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3498–3511, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.305. URL <https://aclanthology.org/2021.eacl-main.305>.
- [18] Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, 2021.
- [19] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 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, 2022.
- [20] Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiase Chen, Hao Zhou, and Lei Li. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.192. URL <https://aclanthology.org/2022.findings-naacl.192>.

- [21] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*, March 2020. URL <http://arxiv.org/abs/2003.10555>. arXiv: 2003.10555.
- [22] Alexis Conneau and Guillaume Lample. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html>.
- [23] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.
- [24] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*, April 2020. URL <http://arxiv.org/abs/1911.02116>. arXiv: 1911.02116.
- [25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- [26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL <http://arxiv.org/abs/1810.04805>. arXiv: 1810.04805.
- [27] Yi Feng, Chuanyi Li, and Vincent Ng. Legal judgment prediction: A survey of the state of the art. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5461–5469. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/765. URL <https://doi.org/10.24963/ijcai.2022/765>. Survey Track.
- [28] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- [29] Claire Grover, Ben Hachey, and Ian Hughson. The HOLJ corpus. supporting summarisation of legal texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 47–54, Geneva, Switzerland, aug 29 2004. COLING. URL <https://aclanthology.org/W04-1907>.
- [30] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct 2021. ISSN 2691-1957. doi: 10.1145/3458754. URL <https://doi.org/10.1145/3458754>.
- [31] Neel Guha, Daniel E. Ho, Julian Nyarko, and Christopher Ré. LegalBench: Prototyping a Collaborative Benchmark for Legal Reasoning, September 2022. URL <http://arxiv.org/abs/2209.06120>. arXiv:2209.06120 [cs].
- [32] Ben Hachey and Claire Grover. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345, 2006. ISSN 1572-8382. doi: 10.1007/s10506-007-9039-z. URL <https://link.springer.com/article/10.1007/s10506-007-9039-z>.
- [33] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [cs]*, December 2021. URL <http://arxiv.org/abs/2111.09543>. arXiv: 2111.09543.

- [34] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*, October 2021. URL <http://arxiv.org/abs/2006.03654>. arXiv: 2006.03654.
- [35] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652, 2017.
- [36] Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, July 2022. URL <http://arxiv.org/abs/2207.00220>. arXiv:2207.00220 [cs].
- [37] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021. URL <http://arxiv.org/abs/2009.03300>. arXiv:2009.03300 [cs].
- [38] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review, 2021.
- [39] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, September 2020. URL <http://arxiv.org/abs/2003.11080>. arXiv:2003.11080 [cs].
- [40] Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D’Orazio. ConflIBERT: A Pre-trained Language Model for Political Conflict and Violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5469–5482. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.400. URL <https://aclanthology.org/2022.naacl-main.400>.
- [41] Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension, December 2022. URL <http://arxiv.org/abs/2212.08204>. arXiv:2212.08204 [cs].
- [42] Allen H. Huang, Amy Y. Zang, and Rong Zheng. Evidence on the information content of text in analyst reports. *ERN: Econometric Modeling in Financial Economics (Topic)*, 2014.
- [43] G Thomas Hudson and Noura Al Moubayed. Muld: The multitask long document benchmark. *arXiv preprint arXiv:2202.07362*, 2022.
- [44] Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction, October 2022. URL <http://arxiv.org/abs/2206.05224>. arXiv:2206.05224 [cs].
- [45] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388, 2021. ISSN 1574-0137. doi: <https://doi.org/10.1016/j.cosrev.2021.100388>. URL <https://www.sciencedirect.com/science/article/pii/S1574013721000289>.
- [46] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL <https://europepmc.org/articles/PMC4878278>.
- [47] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II au2. Natural language processing in the legal domain, 2023.
- [48] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. Natural Language Processing in the Legal Domain, January 2023. URL <https://papers.ssrn.com/abstract=4336224>.

- [49] Kornraphop Kawintiranon and Lisa Singh. PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367. European Language Resources Association, 2022. URL <https://aclanthology.org/2022.lrec-1.801>.
- [50] Mi-Young Kim, Ying Xu, and Randy Goebel. Summarization of legal texts with high cohesion and automatic compression rate. In Yoichi Motomura, Alastair Butler, and Daisuke Bekki, editors, *New Frontiers in Artificial Intelligence*, pages 190–204, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39931-2.
- [51] Anastassia Kornilova and Vladimir Eidelman. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, 2019.
- [52] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*, February 2020. URL <http://arxiv.org/abs/1909.11942>. arXiv:1909.11942.
- [53] Dawn Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. Neural approaches to multilingual information retrieval, 2023.
- [54] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL <https://doi.org/10.1093/bioinformatics/btz682>.
- [55] Johannes Leveling. On the effect of stopword removal for sms-based faq retrieval. In *International Conference on Applications of Natural Language to Data Bases*, 2012.
- [56] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.
- [57] Quanzhi Li and Qiong Zhang. Court opinion generation from case fact description with legal basis. In *AAAI Conference on Artificial Intelligence*, 2021.
- [58] Quanzhi Li and Qiong Zhang. Court opinion generation from case fact description with legal basis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14840–14848, May 2021. doi: 10.1609/aaai.v35i17.17742. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17742>.
- [59] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekogonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, November 2022. URL <http://arxiv.org/abs/2211.09110>. arXiv:2211.09110 [cs].
- [60] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- [61] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. URL <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.

- [62] Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.313. URL <https://aclanthology.org/2021.acl-long.313>.
- [63] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyy Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, 65(4): 782–796, apr 2014. ISSN 2330-1635. doi: 10.1002/asi.23062. URL <https://doi.org/10.1002/asi.23062>.
- [64] Mercedes Martínez-González, Pablo de la Fuente, and Dámaso-Javier Vicente. Reference Extraction and Resolution for Legal Texts. In Sankar K. Pal, Sanghamitra Bandyopadhyay, and Sambhunath Biswas, editors, *Pattern Recognition and Machine Intelligence*, Lecture Notes in Computer Science, pages 218–221, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-32420-1. doi: 10.1007/11590316\_29.
- [65] Maria Medvedeva, Michel Vols, and Martijn Wieling. Judicial decisions of the European Court of Human Rights: looking into the crystal ball. *Proceedings of the Conference on Empirical Legal Studies in Europe 2018*, 2018. URL <https://research.rug.nl/en/publications/judicial-decisions-of-the-european-court-of-human-rights-looking->.
- [66] Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. Dynamic Context Extraction for Citation Classification. 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 539–549, Online only, November 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.aacl-main.41>.
- [67] Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. Benchmarking for biomedical natural language processing tasks with a domain specific bert. *BMC bioinformatics*, 23(1):1–15, 2022.
- [68] Joel Niklaus and Daniele Giofré. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch?, November 2022. URL <http://arxiv.org/abs/2211.17135> [cs].
- [69] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.nllp-1.3>.
- [70] Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 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, November 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.aacl-main.3>.
- [71] Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. Lextreme: A multi-lingual and multi-task benchmark for the legal domain, 2023.
- [72] Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. MultiLegalPile: A 689GB Multilingual Legal Corpus, June 2023. URL <http://arxiv.org/abs/2306.02069>. arXiv:2306.02069 [cs].
- [73] 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. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.

- [74] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://doi.org/10.3115/1073083.1073135>.
- [75] Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents, 2021.
- [76] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In *BioNLP@ACL*, pages 58–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-5006.
- [77] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019. URL <http://arxiv.org/abs/1906.05474>.
- [78] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 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, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL <https://aclanthology.org/2021.emnlp-main.800>.
- [79] Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the Curse of Multilinguality by Pre-training Modular Transformers, May 2022. URL <http://arxiv.org/abs/2205.06266>. arXiv:2205.06266 [cs].
- [80] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. ISSN 1533-7928. URL <http://jmlr.org/papers/v21/20-074.html>.
- [81] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [cs]*, August 2019. URL <http://arxiv.org/abs/1908.10084>. arXiv: 1908.10084.
- [82] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL <http://dx.doi.org/10.1561/1500000019>.
- [83] Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Pantelev, and Partha Talukdar. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages, 2023.
- [84] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, February 2020. URL <http://arxiv.org/abs/1910.01108>. arXiv: 1910.01108.
- [85] 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 Froberg, 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 Névéal, 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, Ruochoen 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 Burynek, 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 Perrián, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, 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 Sängler, 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, Tan-

- may 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. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, November 2022. URL <http://arxiv.org/abs/2211.05100>. arXiv:2211.05100 [cs].
- [86] Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 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), December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.nllp-1.3>.
- [87] Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. WHEN FLUE MEETS FLANG: Benchmarks and large pre-trained language model for financial domain, 2022. URL <http://arxiv.org/abs/2211.00083>.
- [88] Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.
- [89] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities, July 2022. URL <http://arxiv.org/abs/2206.10883>. arXiv:2206.10883 [cs].
- [90] Jerrold Soh, How Khang Lim, and Ian Ernst Chai. Legal Area Classification: A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2208. URL <http://aclweb.org/anthology/W19-2208>.
- [91] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askeel, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubakaran, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack

Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqi, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Śwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimeo Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdah Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Mishnerghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2022. URL <http://arxiv.org/abs/2206.04615>. arXiv:2206.04615 [cs, stat].

- [92] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A Large Language Model for Science, November 2022. URL <http://arxiv.org/abs/2211.09085>. arXiv:2211.09085 [cs, stat].
- [93] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models,

October 2021. URL <http://arxiv.org/abs/2104.08663>. arXiv:2104.08663 [cs].

- [94] 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. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL <https://arxiv.org/abs/2302.13971v1>.
- [95] Jannis Vamvas, Johannes Graën, and Rico Sennrich. Swissbert: The multilingual language model for switzerland. *arXiv e-prints*, pages arXiv–2303, 2023.
- [96] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL <https://aclanthology.org/W18-5446>.
- [97] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. page 30, 2019.
- [98] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- [99] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [100] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v30/Wang13.html>.
- [101] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL <https://arxiv.org/abs/2109.01652>.
- [102] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 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, May 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.eacl-main.217>.
- [103] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance, 2023. URL <http://arxiv.org/abs/2303.17564>.
- [104] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*, July 2018. URL <http://arxiv.org/abs/1807.02478>. arXiv: 1807.02478.
- [105] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84, January 2021. ISSN 2666-6510. doi: 10.1016/j.aiopen.2021.06.003. URL <https://www.sciencedirect.com/science/article/pii/S2666651021000176>.

- [106] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaowei Hua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *COLING*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.419. URL <https://aclanthology.org/2020.coling-main.419>.
- [107] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*, March 2021. URL <http://arxiv.org/abs/2010.11934>. arXiv: 2010.11934.
- [108] Yi Yang, Mark Christopher Siy UY, and Allen Huang. FinBERT: A pretrained language model for financial communications, 2020. URL <http://arxiv.org/abs/2006.08097>.
- [109] Jingtao Zhan, Jiabin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 1503–1512, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462880. URL <https://doi.org/10.1145/3404835.3462880>.
- [110] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv:1912.08777 [cs]*, July 2020. URL <http://arxiv.org/abs/1912.08777>. arXiv: 1912.08777.
- [111] Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *ACL*, pages 7888–7915, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.544. URL <https://aclanthology.org/2022.acl-long.544>.
- [112] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*, February 2020. URL <http://arxiv.org/abs/1904.09675>. arXiv: 1904.09675.
- [113] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. *arXiv:2104.08671 [cs]*, July 2021. URL <http://arxiv.org/abs/2104.08671>. arXiv: 2104.08671 version: 3.
- [114] Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, Yu-Cheng Zhou, and Jia-Rui Lin. Pretrained domain-specific language model for natural language processing tasks in the aec domain. *Comput. Ind.*, 142(C), nov 2022. ISSN 0166-3615. doi: 10.1016/j.compind.2022.103733. URL <https://doi.org/10.1016/j.compind.2022.103733>.
- [115] Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216, 03 2022. ISSN 1527-974X. doi: 10.1093/jamia/ocac040. URL <https://doi.org/10.1093/jamia/ocac040>.
- [116] Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 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, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6\_092. URL [https://doi.org/10.26615/978-954-452-049-6\\_092](https://doi.org/10.26615/978-954-452-049-6_092).

## Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to **[Yes]**, **[No]**, or **[N/A]**. You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? **[Yes]** See Section 1
- Did you include the license to the code and datasets? **[No]** The code and the data are proprietary.
- Did you include the license to the code and datasets? **[N/A]**

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
  - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[Yes]**
  - (b) Did you describe the limitations of your work? **[Yes]** Appendix C
  - (c) Did you discuss any potential negative societal impacts of your work? **[Yes]** Appendix E
  - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[Yes]**
2. If you are including theoretical results...
  - (a) Did you state the full set of assumptions of all theoretical results? **[N/A]**
  - (b) Did you include complete proofs of all theoretical results? **[N/A]**
3. If you ran experiments...
  - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[Yes]** Sections 1, 4, and 5
  - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[Yes]** Appendix G.3
  - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[Yes]** We ran all the text classification experiments with three random seeds each and show standard deviations in the appendix.
  - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]** Appendix G.2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
  - (a) If your work uses existing assets, did you cite the creators? **[Yes]** Sections 1, 2, 4, 5
  - (b) Did you mention the license of the assets? **[Yes]** Section 1
  - (c) Did you include any new assets either in the supplemental material or as a URL? **[Yes]** Sections 1 and 5
  - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[N/A]** The data is already public and not copyrighted.
  - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[Yes]** Section 4
5. If you used crowdsourcing or conducted research with human subjects...
  - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[N/A]**
  - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[N/A]**
  - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[N/A]**

## A Appendix Table of Contents

We include the following supplementary sections as an addition to the main paper:

- B: Access to the Provided Resources
- C: Limitations
- D: Directions of Future Research
- E: Broader Impact
- F: Additional Related Work
- G: More Detailed Experimental Setup
- H: Datasets
- I: Results
- J: Example Generations

## B Access to the Provided Resources

In this section, we provide the URLs to the data, models and code.

### B.1 Data

- Judgment Prediction: [https://huggingface.co/datasets/rcds/swiss\\_judgment\\_prediction\\_xl](https://huggingface.co/datasets/rcds/swiss_judgment_prediction_xl)
- Law Area Prediction: [https://huggingface.co/datasets/rcds/swiss\\_law\\_area\\_prediction](https://huggingface.co/datasets/rcds/swiss_law_area_prediction)
- Criticality Prediction: [https://huggingface.co/datasets/rcds/swiss\\_criticality\\_prediction](https://huggingface.co/datasets/rcds/swiss_criticality_prediction)
- Court View Generation: [https://huggingface.co/datasets/rcds/swiss\\_court\\_view\\_generation](https://huggingface.co/datasets/rcds/swiss_court_view_generation)
- Leading Decision Summarization: [https://huggingface.co/datasets/rcds/swiss\\_leading\\_decision\\_summarization](https://huggingface.co/datasets/rcds/swiss_leading_decision_summarization)
- Information Retrieval: [https://huggingface.co/datasets/rcds/swiss\\_doc2doc\\_ir](https://huggingface.co/datasets/rcds/swiss_doc2doc_ir)
- Citation Extraction: [https://huggingface.co/datasets/rcds/swiss\\_citation\\_extraction](https://huggingface.co/datasets/rcds/swiss_citation_extraction)
- Rulings: [https://huggingface.co/datasets/rcds/swiss\\_rulings](https://huggingface.co/datasets/rcds/swiss_rulings)
- Legislation: [https://huggingface.co/datasets/rcds/swiss\\_legislation](https://huggingface.co/datasets/rcds/swiss_legislation)
- Leading Decisions: [https://huggingface.co/datasets/rcds/swiss\\_leading\\_decisions](https://huggingface.co/datasets/rcds/swiss_leading_decisions)

### B.2 Models

- Legal-CH-RoBERTa<sub>Base</sub>: <https://huggingface.co/joelito/legal-swiss-roberta-base>
- Legal-CH-RoBERTa<sub>Large</sub>: <https://huggingface.co/joelito/legal-swiss-roberta-large>
- Legal-CH-Longformer<sub>Base</sub>: <https://huggingface.co/joelito/legal-swiss-longformer-base>
- Citation Extraction: [https://huggingface.co/rcds/MiniLM-swiss\\_citation\\_extraction-de-fr-it](https://huggingface.co/rcds/MiniLM-swiss_citation_extraction-de-fr-it)

### B.3 Code

- Data Preparation (SwissCourtRulingCorpus): <https://github.com/JoelNiklaus/SwissCourtRulingCorpus>
- Text Classification Experiments (LEXTREME): <https://github.com/JoelNiklaus/LEXTREME>
- Text Generation Experiments: [https://github.com/vr18sub/court\\_view\\_generation](https://github.com/vr18sub/court_view_generation)
- Information Retrieval Experiments (BEIR): <https://github.com/Stern5497/Doc2docBeirIR>
- Citation Extraction Experiments (LEXTREME): <https://github.com/JoelNiklaus/LEXTREME>

## C Limitations

### C.1 General

The research area for language models and benchmarks continues to evolve, and while there is palpable enthusiasm in the field, it is critical to maintain a balanced perspective. Studies, including Bender and Koller [8], have shed light on the limitations of language models and benchmarks, stressing that language models do not truly "learn" meaning and that communities often focus on limited datasets, some of which are borrowed from other fields.

### C.2 Models

Even though English models are plenty, LLMs pre-trained multilingually are very rare. To the best of our knowledge, mT5 is the only multilingual model with variants over 1B parameters, covering German, French and Italian (BLOOM does not contain German and Italian). Additionally, we were limited by very large sequence lengths. We did not have the resources available to run mT5<sub>XL</sub> or mT5<sub>XXL</sub> with sequence lengths over 1K.

### C.3 Data

The process of selecting tasks for benchmarks is typically influenced by the interests of the community or the convenience of available resources, rather than being informed by all-encompassing theories. These constraints present difficulties when trying to explore a model's broader applicability or its capacity for understanding. The data employed in benchmarks is often tied to a specific context and is naturally susceptible to inherent biases. Furthermore, the content of such data may vary significantly from real-world data, is de-contextualized and the uniformity of the task formats may not adequately reflect the diversity of human activities. Regarding our specific context, it is crucial to acknowledge that we cannot generalize Swiss legal data to other countries or different legal systems.

### C.4 Labels

Annotating high-quality datasets is very expensive, especially when experts are needed, such as in the legal domain. Because of our limited budget and in order to arrive at a large amount of labels, we algorithmically generated labels based on metadata information present in the corpus. This metadata is of high quality, being provided by the courts themselves.

### C.5 Text Classification

**Judgment Prediction** While the judgment prediction task is arguably very interesting and also very challenging, it is unlikely to be deployed in practice anytime soon. Ideally, we would want the complaints as input (similar to Semo et al. [86]) instead of the facts description, since this is written by the court itself in part to justify its reasoning. Unfortunately, the complaints are not public in Switzerland, making us rely on the widely available facts description as a close proxy.

**Law Area Prediction** We used information about the chambers at the courts to determine the law areas. Predicting the main law area is not challenging for current models, leading to very high results and thus rendering this task unsuitable for a benchmark. Unfortunately, most chambers cover multiple sub areas, thus ruling them out for the sub area prediction task and considerably reducing dataset size. In conclusion, while this task is very useful in practice for routing requests to the different chambers inside a court, it is relatively unsuitable for a challenging benchmark.

**Criticality Prediction** 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 (BGE-label) and how often this leading decision was cited (Citation-label), we were able to create labels semi-automatically. While we discussed this with lawyers at length and implemented the solution we agreed on finally, this task remains somewhat artificial.



## C.6 Text Generation

**Court View Generation** Court View Generation is an extremely challenging task and thus very well suitable for a benchmark. Current multilingual transformer-based models do not allow processing text in the tens of thousands of tokens. As a consequence, we were forced to look at a simplified version of this task, only considering the facts as input and ignoring relevant case law and legislation. Additionally, we were only able to generate the first 512 tokens of the considerations. We thus invite the community to develop new methods potentially capable of tackling harder versions of this task.

**Leading Decision Summarization** Due to limited resources, we limited our evaluation to mT5<sub>Small/Base/Large</sub>. Future work may investigate large multilingually pretrained generative models on this task. Additionally, one may want to conduct human evaluation on the generated summaries. Finally, we only considered the simple version of this task where we only generate a text-based output. Future work may treat the first and second parts of the summary as extreme multilabel classification problems of relevant citations and relevant keywords from the thesaurus Jurivoc respectively, possibly increasing performance.

## C.7 Citation Extraction

Even though the citations are annotated by the FSCS, we encountered citations that were not marked. However, models achieved very high scores anyway, leading us to exclude it from the benchmark. Future work may investigate this in more detail.

## C.8 Information Retrieval

The labels are constructed with the citations from the considerations. Due to most legal analyses being private, our corpus is restricted to case law and legislation. Constructing a ranking of relevant documents is challenging due to missing information, and thus probably requiring extensive human annotation. Additionally, S-BERT models are usually limited to 512 tokens, being a constraint for this task due to our long documents.

## D Directions for Future Research

The political parties of the judges in the ruling determine in what direction the ruling will go. In future work we would like to enrich the dataset with this information to make models more accurate in the judgment prediction task.

For simplicity, we treat the summary (regeste) of the leading decisions as just one string. Actually, it is composed of important citations in the first part, keywords from a Thesaurus in the second part and a text-based summary in the third part. The first two tasks could be framed as classification or retrieval tasks, possibly improving model performance.

Due to limited context width, we only considered the facts as input to the court view generation task. However, judges and clerks do not only look at the facts when drafting a decision. They consider a myriad of information including possible lower court decisions and relevant case law, legislation and legal analyses. This information is available in our dataset. In the future, we would like to develop systems that are capable of integrating all this information to write the legal reasoning.

Future work may investigate the more difficult Citation Prediction task in addition to the Citation Extraction task. In Citation Prediction, the model only gets the context up to the citation as input and is tasked to predict the citation. This may help lawyers in drafting their texts.

## E Broader Impact

While our research has several positive applications, it is important to acknowledge potential negative societal impacts. Large Language Models (LLMs) and their applications in the legal domain could potentially automate certain tasks traditionally performed by legal professionals, such as legal IR and LDS. While our goal is to support lawyers, it could impact the job market for legal professionals.

Moreover, LLMs can sometimes produce outputs that are not entirely accurate, which can have severe implications when it comes to the legal domain, where precision and factual accuracy are paramount. This could potentially lead to misinformation or misinterpretation of legal texts, impacting legal proceedings and decisions.

While the benchmark focuses on the Swiss legal system, it is important to recognize that law systems are highly culturally and contextually dependent. The understanding and interpretations of legal texts by these models, especially in a multilingual context, might not accurately reflect the nuanced cultural aspects of different regions. This could potentially lead to misrepresentations or misinterpretations, particularly when applied to other legal systems.

Finally, like any AI model, LLMs could be misused to create misinformation or misleading content at scale, especially in languages and domains where automated content generation is still a novel concept. It is crucial to develop and implement robust ethical guidelines and policies to mitigate these risks.

Therefore, while the new benchmark presents exciting opportunities to improve LLMs, it is essential to carefully consider the implications of its use and manage the associated risks effectively. The developers and users of such technology should adhere to ethical guidelines to ensure its responsible use.

## **F Additional Related Work**

### **F.1 Domain-Specific Pretraining**

General-purpose language models are trained on generic text corpora such as Wikipedia and evaluated on widely used benchmarks such as GLUE [96]. However, domain-specific models need focused datasets for training and specialized benchmarks for assessing the quality of the model. The following examples illustrate the increase in performance when using domain-specific datasets and benchmarks.

In the biomedical area of natural language processing (BioNLP), Lee et al. [54] created for the first time a domain-specific LM based on BERT [26] by pre-training it on biomedical text corpora. They used PubMed abstracts (4.5B words) and PubMed Central (PMC) full-text articles (13.5B words). The resulting domain-specific LM BioBERT achieved higher F1 scores than BERT in the biomedical NLP tasks named entity recognition (0.62) and relation extraction (2.80), and a higher mean reciprocal rank (MRR) score (12.24) in the biomedical question-answering task. In 2022, those scores were outperformed. Naseem et al. [67] conducted a domain-specific pre-training of ALBERT [52] using only text from the biomedical field (PubMed) and from the "Medical Information Mart for Intensive Care" (MIMIC-III), a large, de-identified and publicly-available collection of medical records [46]. One domain-specific benchmark applied to test BioALBERT originates from Gu et al. [30] who created BLURB, the Biomedical Language Understanding and Reasoning Benchmark. Naseem et al. [67] found that BioALBERT exceeded the state-of-the-art models by 11.09% in terms of micro averaged F1-score (BLURB score). Another biomedical NLP benchmark is BLUE, the "Biomedical Language Understanding Evaluation" [77]. It covers five tasks (sentence similarity, named entity recognition, relation extraction, document classification, inference) with ten datasets from the biomedical and clinical area. BioALBERT also includes all datasets and tests from BLUE thus presenting the most comprehensive domain-specific model and benchmark in the biomedical area at the moment.

In the financial domain, FinBERT was pretrained 2020 by Yang et al. [108] using financial data. The text corpora consisted of 203'112 corporate reports (annual and quarterly reports from the Securities Exchange Commission SEC), 136'578 earnings call transcripts (conference call transcripts from CEOs and CFOs), and 488'494 analyst reports (textual analysis of the company) resulting in 3.3B tokens. For testing FinBERT, Yang et al. [108] used the Financial Phrase Bank dataset with 4'840 sentiment classifications [63], the AnalystTone Dataset with 10'000 sentences [42], and FiQA Dataset with 1'111 sentences from an open challenge dataset for financial sentiment analysis (Financial Opinion Mining and Question Answering). The results show that the domain-specific FinBERT outperforms the generic BERT models in all of these financial datasets. An improved financial domain LM was released 2022 by [87] by introducing FLANG-BERT, the Financial LANGUAGE Model. They also created a domain-specific benchmark, Financial Language Understanding Evaluation (FLUE). Recently in May 2023, Bloomberg announced the BloombergGPT model, a Large Language Model

(LLM) for the financial domain [103]. However, next to some experience on the training process no datasets, benchmarks, or weights have been released publicly.

Numerous other domain-specific LMs have been created since the rise of BERT. They all outperform general-purpose LMs. For instance, SciBERT is a pretrained LM based on scientific publications and evaluated on a suite of tasks in different scientific domains [6]. ConflBERT is built to improve monitoring of political violence and conflicts [40] and PoliBERTweet is used to analyze political content on Twitter [49]. Cybersecurity is another important area thus [1] pretrained a LM on a large corpus of cybersecurity text. To improve IR tasks in the architecture, engineering, and construction (AEC) industry, Zheng et al. [114] pretrained BERT on a corpus of regulatory text. Also, the domain-specific model BlueBERT [77] from the biomedical domain has been further pretrained and evaluated on more narrow, cancer-related vocabularies, resulting in CancerBERT [115].

In the legal domain Chalkidis et al. [15] pretrained LegalBERT on EU and UK legislation, ECHR and US cases and US contracts. Zheng et al. [113] pretrained CaseHoldBERT on US caselaw. Henderson et al. [36] trained PoL-BERT on their 256 GB diverse Pile of Law corpus. Niklaus and Giofr  [68] pretrained longformer models using the Replaced Token Detection (RTD) task [21] on the Pile of Law. Hua et al. [41] pretrained reformer models with RTD on 6 GB of US caselaw. Finally, Niklaus et al. [72] released a large multilingual legal corpus and trained various legal models on it.

## F.2 Judgment Prediction

The domain of Legal Judgment Prediction (LJP) centers around the crucial task of predicting legal case outcomes given the provided facts. In the landscape of LJP research, there have been significant advances focusing on diverse languages, jurisdictions, and input types. Researchers have utilized a variety of datasets, each with their unique characteristics and annotations, to analyze and predict case outcomes [27, 2, 116, 65, 13].

In the context of Chinese criminal cases, notable efforts have been made by Xiao et al. [104, 105], where they utilized the CAIL2018 dataset, which consists of over 2.6M cases and provides annotations for Law Article, Charge, and Prison Term, among others.

Focusing on the Indian and Swiss jurisdictions, Malik et al. [62] and Niklaus et al. [69, 70] employed the ILDC and SJP datasets respectively, both using binary labels. The ILDC dataset, with over 34K Indian Supreme Court cases, offers sentence-level explanations along with Court Decision annotations, while the SJP corpus is trilingual, containing judgments from Switzerland in German, French, and Italian, and provides annotations like the publication year, legal area, and the canton of origin.

European jurisdictions have been explored using the ECHR2019 and ECHR2021 datasets [13, 18]. These corpora feature cases from the European Court of Human Rights, annotated for Violation, Law Article, and Alleged Law Article, among others, with the latter also providing paragraph-level rationales.

The FCCR dataset, containing over 126K cases from France, has been used to predict Court Decisions with different setups, offering additional annotations such as the date of the court ruling and the law area [116].

Recently, Semo et al. [86] introduced a new perspective on LJP, applying it to US class action cases. The proposed task involves predicting the judgment outcome based on the plaintiff’s pleas, further expanding the scope of LJP research and making the task more realistic.

These efforts underscore the breadth of LJP research, demonstrating its applicability across multiple jurisdictions, languages, and legal systems, and its potential in assisting legal professionals and enhancing access to justice.

## F.3 Criticality Prediction

Chalkidis et al. [13] introduced the Importance Prediction task, which predicts the importance of a ECtHR case on a scale from 1 (key case) to 4 (unimportant). Legal experts defined and assigned these labels for each case, representing a significant contrast to our approach where labels were algorithmically determined. This is to our knowledge the only comparable task to Criticality Prediction.

#### **F.4 Law Area Prediction**

Although not widely studied, several notable works have focused on LAP. Şulea et al. [116] worked on the Law Judgment Prediction (LJP) task, using a dataset of over 126K cases from the French Supreme Court. The study used Linear Support Vector Machines (SVMs) to classify cases into one of eight law areas, using the entire case description as input. This approach yielded an F1 score of 90%. Soh et al. [90] conducted a similar study using a dataset of 6K judgments in English from the Singapore Supreme Court. These judgments were mapped into 30 law areas. Several text classifiers were used in the study, achieving a macro F1 score of up to 63.2%.

#### **F.5 Court View Generation**

Over the past decade, text generation in the field of Legal NLP has been underexplored [48], especially in comparison to tasks such as classification and information extraction. Li and Zhang [58] utilize Chinese case facts, as well as charge (formal accusations of crimes) and law article information, to generate court opinions. A key difference from our task is the shortness of their opinions (avg. 31/34 tokens), while ours span approximately four thousand tokens on average. With the emergence of powerful generative models, we expect a surge in research activity in this area, necessitating challenging benchmarks to assess progress effectively.

#### **F.6 Leading Decisions Summarization**

In the field of legal text summarization, several noteworthy contributions have been made [29, 32, 50, 45], with the BillSum [51] and Multi-LexSum [89] datasets being particularly significant. The creators of the BillSum dataset focused on summarizing 22K bills from the US Congress and the state of California. They also applied transfer learning in summarization from federal to state laws. Models based on BERT and TF-IDF, as well as a combination of both, have been evaluated on this dataset. The BillSum dataset focuses on English language documents related to the US legislative environment. The Multi-LexSum dataset is another significant development in the area of legal text summarization. It targets long civil rights lawsuits, with an average length of over 75K words. This 9K-document dataset allows for in-depth study at different summary lengths: short (25 words), medium (130 words), and long (650 words), a unique feature of the Multi-LexSum dataset. Models based on BART [56] and PEGASUS [110] were evaluated on this dataset. Like BillSum, the Multi-LexSum dataset is primarily for the English language and is relevant to the US legal setting.

#### **F.7 Citation Extraction**

Early work from Martínez-González et al. [64] extract citations from legal text with patterns. Nambanoor Kunnath et al. [66] studied the effect of differing context size for citation classification in scientific text. Taylor et al. [92] considered the more difficult Citation Prediction task on scientific text and found that larger models are more true to the real citation distribution, whereas smaller models tend to output the most frequent citations most of the times.

#### **F.8 Information Retrieval**

Lawrie et al. [53] revisited the challenges of multilingual IR and proposed neural approaches to address this issue. They demonstrated that combining neural document translation with neural ranking resulted in the best performance in their experiments conducted on the MS MARCO dataset [4]. However, this approach is computationally expensive. To mitigate this issue, they showed that using a pre-trained XLM-R multilingual model to index documents in their native language resulted in only a two percent difference in effectiveness. XLM-R is a transformer-based masked language model that employs self-supervised training techniques for cross-lingual understanding [23]. Lawrie et al. [53] crucially utilized mixed-language batches from the neural translation of MS MARCO passages.

A widely used technique is BM-25, which is an improved retrieval method that considers the term frequencies and takes into account the saturation effect and document length [82]. The saturation effect refers to the point where the relevance of a term stops increasing, even if it appears many times in a document. This issue is mitigated through the use of an additional parameter,  $k$ . Additionally, longer documents are more likely to contain a higher number of occurrences of a term simply because they contain more words, not necessarily because the term is more relevant to the document, which

is why parameter  $b$  is used. The BM-25 score is calculated using the Inverse Document Frequency (IDF), Term Frequency (TF), queries  $Q$ , documents  $d$ , and term  $t$ .

$$BM25(d, Q, b, k) = \sum_{t \in Q} IDF(t) \frac{(k+1)TF(t,d)}{(1-b)(b \cdot A) + TF(t,d)}$$

Chalkidis et al. [17] proposed a new IR task called REG-IR, which deals with longer documents in the corpus and entire documents as queries. This task is an adaptation of Document-to-Document (Doc2Doc) IR, which aims to identify a relevant document for a given document. The authors observed that neural re-rankers underperformed due to contradicting supervision, where similar query-document pairs were labeled with opposite relevance. Additionally, they demonstrated for long documents that using BM25 as a document retriever in a two-stage approach often results in underperformance since the parameters  $k$  and  $b$  are often not optimal when using standard values. The problem of noise filtering of long documents was also addressed by using techniques like stopwords removal. However, as seen in [55], this approach can have a negative effect on performance. The best pre-fetcher for long documents was found to be C-BERTs [16], which are trained on classifying documents using predefined labels.

Thakur et al. [93] proposed a novel evaluation benchmark for IR that encompasses a wide range of approaches, including BM25, dense, and re-ranking models. They found that while BM25 is computationally expensive, it provides a robust baseline, whereas other models failed to achieve comparable performance. Their findings suggest that there is still much room for improvement in this area of NLP. Efficient retrieval of relevant information is crucial for many NLP tasks, and these results highlight the need for continued research in this area.

## G More Detailed Experimental Setup

### G.1 Pretraining Legal Models

(a) We warm-start (initialize) our models from the original XLM-R checkpoints (base or large) of Conneau and Lample [22]. Model recycling is a standard process followed by many [101, 73] to benefit from starting from an available “well-trained” PLM, rather from scratch (random). XLM-R was trained on 2.5TB of cleaned CommonCrawl data in 100 languages.

(b) We train a new tokenizer of 128K BPEs on the training subsets to better cover legal language across languages. However, we reuse the original XLM-R embeddings for all lexically overlapping tokens [78], 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 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. [25]<sup>11</sup>, based on the findings of Wettig et al. [102].

(d) For both training the tokenizer and the our legal models, we use a sentence sampler with exponential smoothing of the sub-corpora sampling rate following Conneau and Lample [22] and Raffel et al. [80], since there is a disparate proportion of tokens across cantons and languages (Figures 7 and 6) and we aim to preserve per-canton and language capacity, i.e., avoid overfitting to the majority (almost 50% of the total number of texts) German texts.

(e) We consider mixed cased models, i.e., both upper- and lowercase letters covered, similar to all recently developed large PLMs [22, 80, 9].

(f) To better account for long contexts often found in legal documents, we continue training the base-size multilingual model on long contexts (4096 tokens) with windowed attention (128 tokens

<sup>11</sup>Devlin et al. [25] – and many other follow-up work – used a 15% masking ratio, and a recipe of 80/10/10% of predictions made across masked/randomly-replaced/original tokens.

window size) [7] for 50K steps, dubbing it Legal-Swiss-LF-base. We use the standard 15% masking probability and increase the learning rate to  $3e-5$  before decaying but otherwise use the same settings as for training the small-context models.

## G.2 Resources Used

The experiments were performed on internal university clusters on NVIDIA GPUs with the following specifications: 24GB RTX3090, 32GB V100, 48GB A6000, and 80GB A100. We used an approximate total of 160, 20, and 2 GPU days for the text classification, text generation and information retrieval experiments.

## G.3 Hyperparameters

**Text Classification** For all models and datasets, a learning rate of  $1e-5$  was used without any tuning. Each experiment was executed with three random seeds (1-3), and the batch size was tailored for each task and corresponding computational resource. If the GPU memory was inadequate, gradient accumulation was employed as a workaround to arrive at a final batch size of 64. The training was conducted with early stopping based on validation loss, maintaining a patience level of 5 epochs. Due to the considerable size of the judgment prediction dataset and the extended duration of the experiment, training was limited to a single epoch with evaluations after every 1000th step. To reduce costs, we utilized AMP mixed precision during the training and evaluation phases whenever it did not lead to overflows (e.g., mDeBERTa-v3). We established the max-sequence-length (determined by the product of max-segment-length and max-segments in the hierarchical setup [2, 69, 70]) based on whether we used Facts: 2048 (128 X 16), or Considerations: 4096 (128 X 32).

**Text Generation** For the main CVG dataset, we trained our models for only one epoch (because of the large training set) with a final batch size of 16, using gradient accumulation as needed. We performed evaluations every 1000 steps. For the smaller origin dataset, we increased the number of epochs to 100 and evaluated every 100 steps. For the LDS task, we adjusted the training to 10 epochs.

**Information Retrieval** For the BM25 model, we used the same parameters as used in the BEIR paper [93], chosen were  $k = 0.9$  and  $b = 0.4$ . For the SBERT model, we were limited to a maximum sequence length of 512 tokens. During training, we used 1 epoch with 5000 evaluation steps. When training with hard negative examples, we used 5 negative examples for each query.

## H Datasets

In this section we provide additional information about the datasets. Table 8 provides additional information about general dataset metadata.

Table 8: Listing of cantons, courts, chambers, law-areas

Metadata	Number	Examples
Cantons	26 (+1)	Aargau (AG), Bern (BE), Basel-Stadt (BS), Solothurn (SO), Ticino (Ti), Vaud (VD),... (+ Federation (CH))
Courts	184	Cantonal Bar Supervisory Authority, Supreme Court, administrative authorities, Tax Appeals Commission, Cantonal Court, Federal Administrative Court, ...
Chambers	456	GR-UPL0-01, AG-VB-002, CH-BGer-011, ZH-OG-001, ZG-VG-004, VS-BZG-009, VD-TC-002, TI-TE-001, ...
Law-Areas	4	Civil, Criminal, Public, Social
Languages	5	German, French, Italian, Romansh, English

### H.1 Rulings

Figures 3 and 4 provide an overview of the distribution of languages and cantons in the rulings dataset respectively. Figure 5 shows the length distribution of the cases.

### H.2 Legislation

Figures 6 and 7 provide an overview of the distribution of languages and cantons in the legislation dataset respectively. Figure 8 shows the length distribution of the legislation texts.

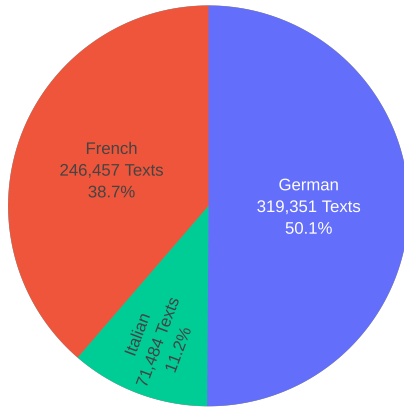


Figure 3: Language distribution of rulings texts

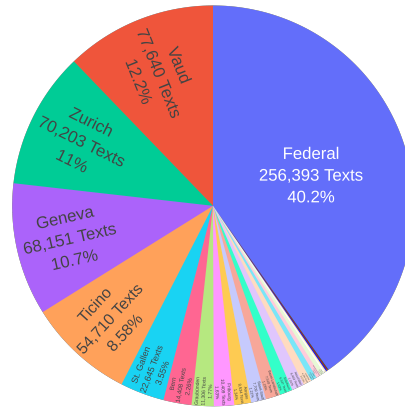


Figure 4: Cantonal distribution of rulings texts

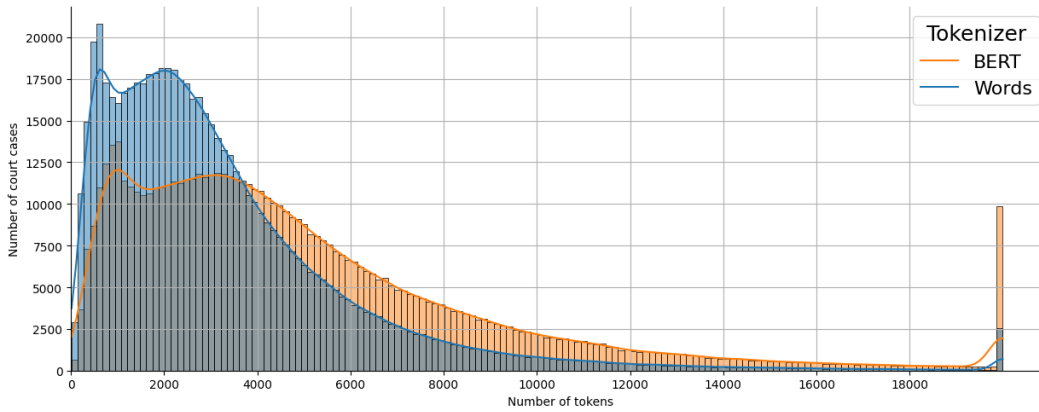


Figure 5: Rulings text length distribution

### H.3 Leading Decisions

Figures 9 and 10 show the length distributions for the facts and considerations of the Leading Decisions dataset.

### H.4 Law Area Prediction

Figures 11 and 12 show the length distributions for the facts of the LAP and SLAP datasets.

### H.5 Criticality Prediction

Figures 13 and 14 show the length distributions for the facts and the considerations of the CP dataset respectively.

### H.6 Judgment Prediction

Figure 15 shows the length distribution for the facts of the JP dataset.

### H.7 Court View Generation

Figures 16 and 17 show the length distributions for the facts and the considerations of the CVG dataset respectively.

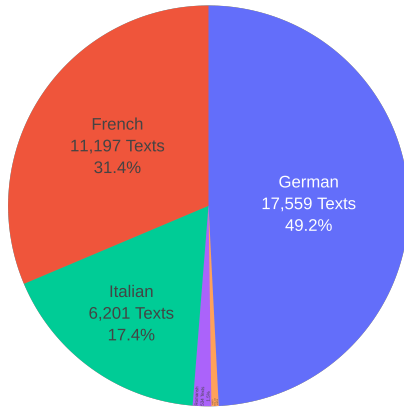


Figure 6: Language distribution of legislation texts

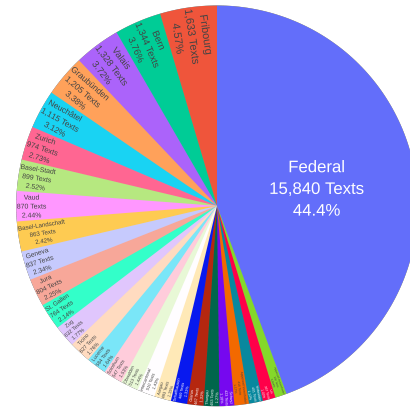


Figure 7: Cantonal distribution of legislation texts

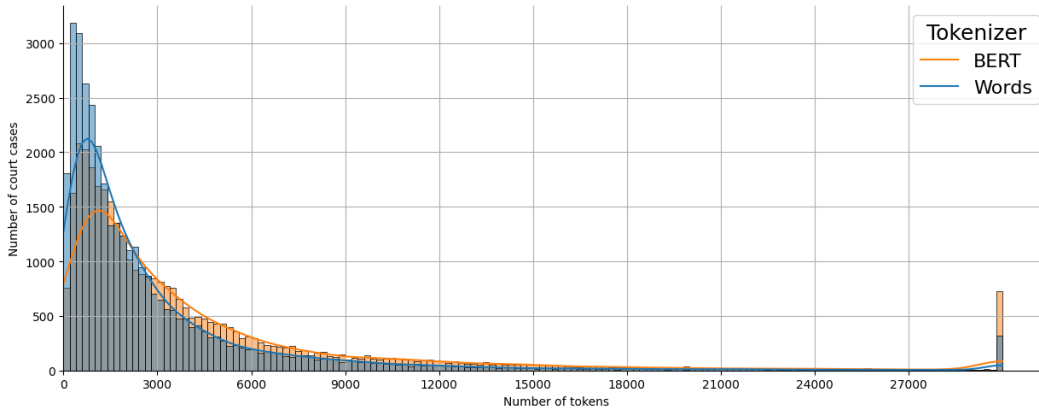


Figure 8: Legislation text length distribution

## H.8 Leading Decision Summarization

Figures 18 and 19 show the length distributions for the input text and the summary of the LDS dataset respectively.

## H.9 Information Retrieval

Figure 20 shows the length distribution for the facts of the IR dataset. Figure 21 shows the structure of the corpus, queries and qrels for the IR task.

# I Results

## I.1 Information Retrieval

Table 9: Results IR: using a subset of 100 queries and only relevant documents in the corpus resulting in an easier task

Model	Additional	$Rcap@1 \uparrow$	$Rcap@10 \uparrow$	$Rcap@100 \uparrow$	$NDCG@1 \uparrow$	$NDCG@10 \uparrow$	$NDCG@100 \uparrow$
Train + Evaluate S-BERT	sbert-legal-xlm-roberta-base	32.32	32.34	81.77	32.32	30.89	49.11
Train + Evaluate S-BERT	sbert-legal-swiss-roberta-base	<b>36.36</b>	<b>35.68</b>	76.03	<b>36.36</b>	<b>34.54</b>	49.90
Train + Evaluate S-BERT	distiluse-base-multilingual-cased-v1	22.22	30.35	84.38	22.22	25.72	48.66
Evaluate S-BERT	distiluse-base-multilingual-cased-v1	8.08	11.83	43.35	8.08	10.55	21.56
Train(HN) + Evaluate S-BERT	distiluse-base-multilingual-cased-v1	27.27	33.94	<b>86.81</b>	27.27	30.09	<b>52.03</b>
Dim Reduction	distiluse-base-multilingual-cased-v1	0.00	1.59	5.43	0.00	1.17	2.41
Cross Encoder	distiluse-base-multilingual-cased-v1	5.94	8.04	14.20	2.97	1.84	7.35
Lexical		5.94	8.04	14.20	5.94	8.52	10.64
ML Lexical	'German'	9.90	8.41	15.19	9.90	9.14	11.58



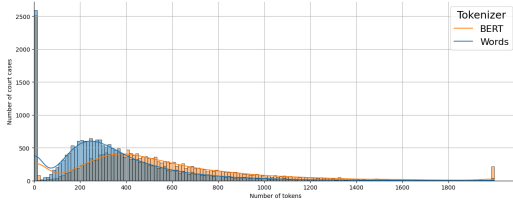


Figure 9: Leading Decisions facts length distribution

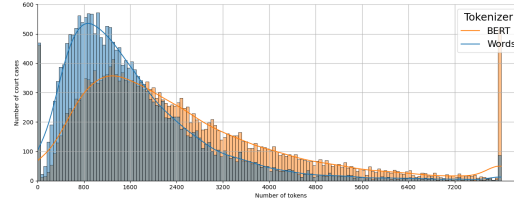


Figure 10: Leading Decisions considerations length distribution

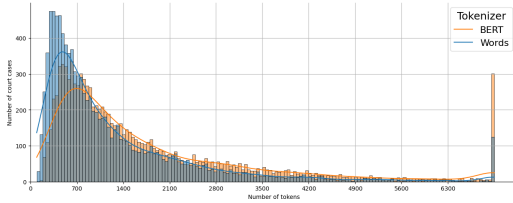


Figure 11: Law Area Prediction facts length distribution

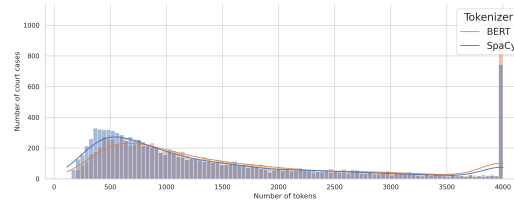


Figure 12: Law Sub Area Prediction facts length distribution

For the ML Lexical Retrieval model a main language must be chosen, indicated with German, French and Italian. Dataset adaptations are indicated with: (S) stopword removal, (SL) using only single language links, (DE/FR/IT) using only queries in one language. Table 9 shows the results of the IR task on a subset of 100 queries and with only relevant documents while Table 10 shows more detailed results using all queries.

## I.2 Court View Generation

Table 11 shows the results of the CVG task from both, the main and the origin dataset.

## I.3 Text Classification

Table 12 shows more detailed results on the text classification datasets including standard deviations across seeds.

Table 13 shows ChatGPT results on the validation sets of the text classification tasks for reference. Note the very low aggregate performance (due to the harmonic mean giving a lot of weight to the low LAP performance). If we exclude LAP, the aggregate score would be 26.9.

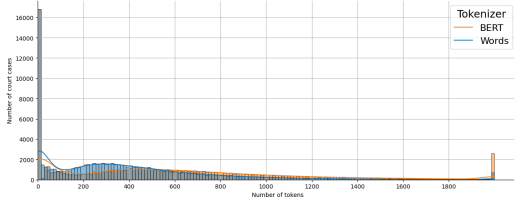


Figure 13: Criticality Prediction facts length distribution

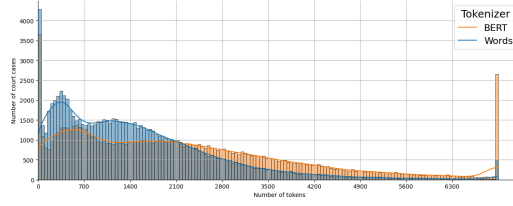


Figure 14: Criticality Prediction considerations length distribution

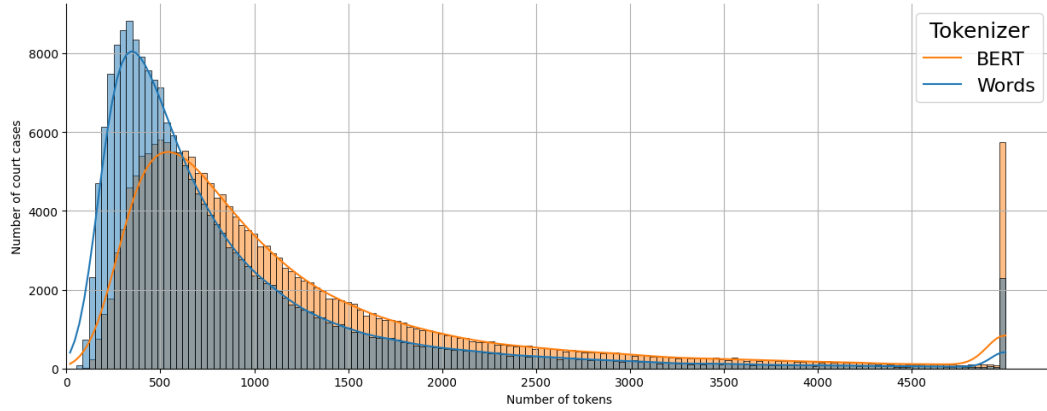


Figure 15: Judgment Prediction facts length distribution

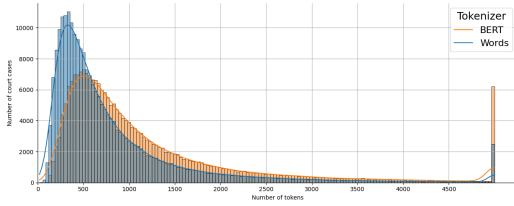


Figure 16: Court View Generation facts length distribution

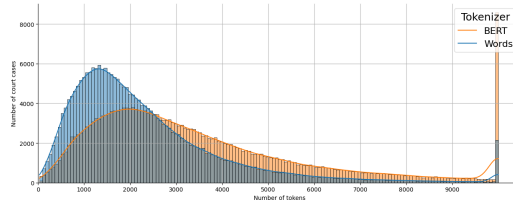


Figure 17: Court View Generation considerations length distribution

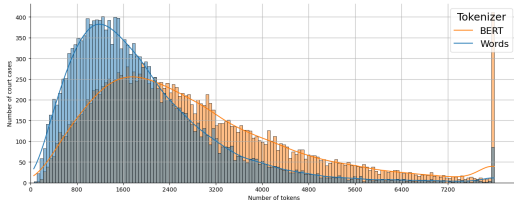


Figure 18: Leading Decision Summarization input length distribution

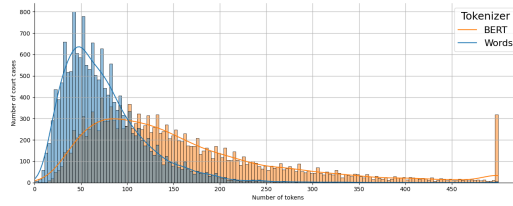


Figure 19: Leading Decision Summarization summary length distribution

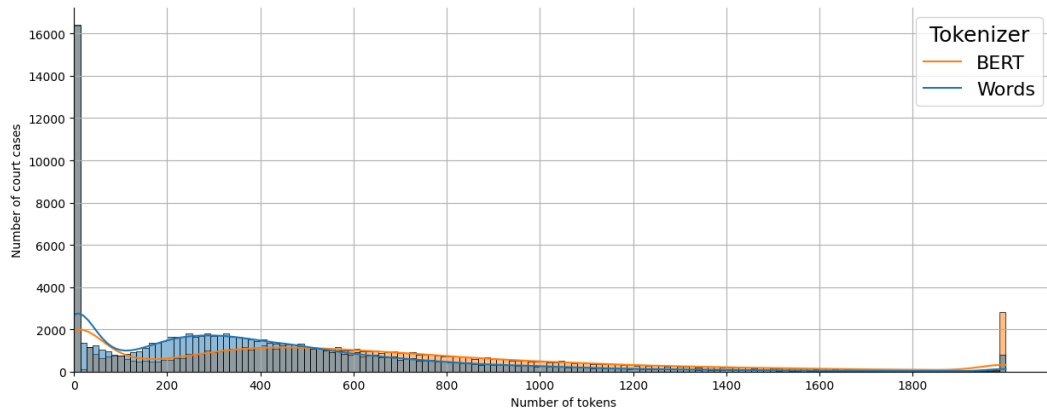


Figure 20: Information Retrieval facts length distribution

```

"corpus": {
  "decision_id_bge": {
    "title": "file number",
    "text": "facts and considerations of a case"
  },
  "law_id": {
    "title": "",
    "text": "text of a law"
  }
},
"queries": {
  "decision_id_bger_1": "facts of a case",
  "decision_id_bger_2": "facts"
},
"qrels": {
  "decision_id_bger_1": {"law-id": 1, "decision-id-bge": 1},
  "decision_id_bger_2": {"law-id": 1}
}

```

Figure 21: Structure of corpus, queries and qrels for IR task

Table 10: Results IR Additional: Results IR Abbreviations: Capped Recall, NDCG, **distiluse-base-multilingual-cased-v1**, joelito/swiss-legal-**roberta-base**, joelito/legal-**xlm-roberta-base**, Train, Hard Negative, Evaluate, **S-Bert**, **Dim** Reduction

Model	Adaption	$R@1 \uparrow$	$R@10 \uparrow$	$R@100 \uparrow$	$N@1 \uparrow$	$N@10 \uparrow$	$N@100 \uparrow$
LR		8.38	6.43	15.76	8.38	6.66	10.23
LR	S	<b>10.64</b>	7.57	16.47	<b>10.64</b>	8.04	11.33
LR	SL	7.91	<b>9.99</b>	<b>32.46</b>	9.13	<b>9.65</b>	<b>18.03</b>
MLR	'German'	8.69	6.54	15.99	8.69	6.82	10.43
MLR	'German'	S <b>10.88</b>	7.65	16.80	<b>10.88</b>	8.14	11.53
MLR	'German'	SL 8.05	<b>9.94</b>	<b>32.63</b>	9.30	<b>9.70</b>	<b>18.17</b>
MLR	'French'	<b>11.37</b>	<b>7.74</b>	<b>16.54</b>	<b>11.37</b>	<b>8.34</b>	<b>11.51</b>
MLR	'French'	S 10.97	7.60	16.52	10.97	8.14	11.41
MLR	'Italian'	10.08	7.118	16.294	10.08	7.582	11.021
MLR	'English'	8.38	6.43	15.76	8.38	6.66	10.23
T+E SB	xlm	2.77	2.58	10.17	6.36	5.66	12.03
T+E SB	rob	3.97	3.47	12.28	9.12	7.76	15.16
E SB	dist	0.90	0.75	2.64	2.06	1.70	3.31
T+E SB	dist	4.4	3.92	12.64	10.11	8.76	16.16
T+E SB	dist	S 4.69	4.14	13.39	10.77	9.27	17.05
T+E SB	dist	SL 1.79	3.92	14.17	4.03	6.17	12.91
SB T(HN)+E	dist	3.97	4.46	13.36	9.12	9.21	16.87
SB T(HN)+E	dist	S 3.76	4.75	12.80	8.64	9.66	16.57
SB T(HN)+E	dist	SL 2.34	4.37	14.43	5.27	6.99	13.75
T+E SB	dist	DE 4.22	4.49	15.21	8.21	8.15	15.86
T+E SB	dist	DE SL 4.06	8.47	29.43	4.51	6.73	13.78
T+E SB	dist	FR 1.88	2.2	9.19	5.77	6.22	13.94
T+E SB	dist	FR SL 2.69	5.68	27.28	3.0	4.59	11.11
T+E SB	dist	IT 0.22	0.24	0.79	5.43	5.74	11.44
T+E SB	dist	IT SL 1.71	4.54	16.24	1.91	3.38	6.83
Dim	dist	0.71	0.62	2.42	1.64	1.4	2.95

Table 11: Results of Court View Generation task. 'In Len' denotes input length in tokens. **Bold**: best within model; underlined: best overall.

Model	In Len $\uparrow$	Main Scores $\uparrow$					Origin Scores $\uparrow$				
		BERT	BLEU	MET	R1 / R2 / RL			BERT	BLEU	MET	R1 / R2 / RL
mT5 <sub>Large</sub>	2048	<b>75.74</b>	<b>66.92</b>	<b>34.44</b>	<b>34.91</b> / <b>15.58</b> / <b>33.53</b>	76.24	<b>62.59</b>	32.25	34.80 / 16.11 / 33.58		
mT5 <sub>Large</sub>	1024	75.56	66.68	34.02	34.26 / 14.72 / 32.87	74.99	58.35	31.06	33.35 / 14.80 / 32.16		
mT5 <sub>Large</sub>	512	75.27	66.12	33.48	33.61 / 14.26 / 32.21	<b>76.33</b>	62.08	<b>32.92</b>	<b>36.61</b> / <b>18.17</b> / <b>34.84</b>		
mT5 <sub>Base</sub>	2048	75.01	65.48	32.89	33.23 / 13.57 / 31.89	75.99	<b>63.39</b>	<b>34.15</b>	36.48 / <b>18.81</b> / 35.58		
mT5 <sub>Base</sub>	1024	<b>75.15</b>	<b>65.73</b>	<b>33.15</b>	<b>33.49</b> / <b>13.96</b> / <b>32.18</b>	76.07	60.99	33.50	<b>37.68</b> / <b>18.79</b> / <b>36.58</b>		
mT5 <sub>Base</sub>	512	74.89	65.55	32.66	32.66 / 13.16 / 31.35	<b>76.08</b>	62.21	32.80	36.40 / 17.58 / 34.98		
mT5 <sub>Small</sub>	2048	<b>74.13</b>	<b>63.97</b>	<b>30.96</b>	<b>31.29</b> / <b>11.01</b> / <b>29.90</b>	75.23	56.59	30.71	34.68 / 13.64 / 33.24		
mT5 <sub>Small</sub>	1024	74.00	63.70	30.68	31.05 / 10.77 / 29.64	<b>75.75</b>	58.99	31.17	34.62 / 14.25 / <b>33.91</b>		
mT5 <sub>Small</sub>	512	73.92	63.83	30.57	30.58 / 10.35 / 29.20	75.63	<b>61.12</b>	<b>32.33</b>	<b>35.16</b> / <b>14.45</b> / 33.72		

Table 12: Configuration aggregate scores with standard deviations.

Model	CPB-F	CPB-C	CPC-F	CPC-C	SLAP-F	SLAP-C	JP-F	JP-C	Agg.
MiniLM	54.7 $\pm$ 1.9	65.8 $\pm$ 1.6	9.8 $\pm$ 2.8	20.8 $\pm$ 3.0	59.7 $\pm$ 3.8	61.1 $\pm$ 3.7	58.1 $\pm$ 0.4	78.5 $\pm$ 2.3	32.4
DistilBERT	56.2 $\pm$ 0.5	65.4 $\pm$ 1.7	19.6 $\pm$ 1.1	22.1 $\pm$ 0.4	63.7 $\pm$ 11.7	65.9 $\pm$ 6.4	59.9 $\pm$ 0.9	75.5 $\pm$ 3.3	42.1
mDeBERTa-v3	55.1 $\pm$ 2.0	69.8 $\pm$ 2.8	21.0 $\pm$ 3.6	17.5 $\pm$ 4.4	63.8 $\pm$ 6.3	59.3 $\pm$ 7.6	60.6 $\pm$ 0.9	77.9 $\pm$ 2.6	40.2
XLM-R <sub>Base</sub>	57.2 $\pm$ 1.5	65.9 $\pm$ 3.2	21.3 $\pm$ 1.5	23.7 $\pm$ 1.9	67.2 $\pm$ 15.9	73.4 $\pm$ 2.5	60.9 $\pm$ 0.6	79.7 $\pm$ 2.5	44.6
XLM-R <sub>Large</sub>	56.4 $\pm$ 1.8	67.9 $\pm$ 1.9	24.4 $\pm$ 7.2	28.5 $\pm$ 2.2	65.1 $\pm$ 8.5	78.9 $\pm$ 4.6	60.8 $\pm$ 0.6	80.9 $\pm$ 2.4	48.4
X-MOD <sub>Base</sub>	56.6 $\pm$ 1.8	67.8 $\pm$ 2.9	20.0 $\pm$ 3.0	20.6 $\pm$ 3.5	63.9 $\pm$ 10.1	64.4 $\pm$ 7.0	60.5 $\pm$ 0.6	79.1 $\pm$ 2.6	41.9
SwissBERT (xlm-vocab)	56.2 $\pm$ 0.7	67.3 $\pm$ 4.7	25.7 $\pm$ 8.3	23.0 $\pm$ 4.0	63.2 $\pm$ 3.7	75.1 $\pm$ 5.2	61.4 $\pm$ 0.6	79.4 $\pm$ 2.5	44.6
Legal-ch-R <sub>Base</sub>	57.6 $\pm$ 2.7	72.8 $\pm$ 2.1	23.1 $\pm$ 2.5	22.5 $\pm$ 6.2	81.6 $\pm$ 2.6	83.0 $\pm$ 11.1	64.0 $\pm$ 1.3	86.4 $\pm$ 1.9	46.9
Legal-ch-R <sub>Large</sub>	57.4 $\pm$ 1.9	70.8 $\pm$ 1.1	21.3 $\pm$ 1.4	23.3 $\pm$ 5.3	80.4 $\pm$ 7.9	84.9 $\pm$ 9.9	62.8 $\pm$ 0.9	87.0 $\pm$ 2.3	46.2
Legal-ch-LF <sub>Base</sub>	58.1 $\pm$ 2.1	70.8 $\pm$ 2.9	21.4 $\pm$ 2.9	17.4 $\pm$ 8.6	80.2 $\pm$ 12.8	83.5 $\pm$ 11.4	65.4 $\pm$ 1.7	86.4 $\pm$ 1.8	42.8

Table 13: Results with ChatGPT (gpt-3.5-turbo) on the validation sets performed on June 7, 2023.

Config	# samples	Macro F1
CPB-F	998	47.0
CPB-C	775	44.3
CPC-F	563	21.3
CPC-C	563	18.7
SLAP-F	911	0.1
SLAP-C	690	0.3
JP-F	946	30.8
JP-C	484	21.7
Agg.		0.6

## J Example Generations

Tables 14 and 15 show excerpts of examples produced by the best model for the CVG and LDS tasks respectively.

Table 14: Nine examples of generated considerations by mT5<sub>Large</sub> with input length 2048 across three languages, showcasing high, average, and low scored outputs in CVG Task

Target	considerations: Erwägungen: 1. 1.1 Der angefochtene Entscheid ist in Anwendung von Sozialversicherungsrecht ergangen. Die Sozialversicherungsrechtliche Abteilung des Verwaltungsgerichts beurteilt gemäss Art. 57 des Bundesgesetzes vom 6. Oktober 2000 über den Allgemeinen Teil des Sozialversicherungsrechts (ATSG; SR 830.1) i.V.m. Art. 54 Abs. 1 lit. a des kantonalen Gesetzes vom 11. Juni 2009 über die Organisation der Gerichtsbehörden und der Staatsanwaltschaft (GSOG; BSG 161.1) Beschwerden gegen solche Entscheide. Die Beschwerdeführerin ist im vorinstanzlichen Verfahren mit ihren Anträgen nicht durchgedrungen, durch den angefochtenen Entscheid berührt und hat ein schutzwürdiges Interesse an dessen Aufhebung, weshalb sie zur Beschwerde befugt ist (Art. 59 ATSG). Die örtliche Zuständigkeit ist gegeben (Art. 58 ATSG). Da auch die Bestimmungen über Frist (Art. 60 ATSG) sowie Form (Art. 61 lit. b ATSG; Art. 81 Abs. 1 i.V.m. Art. 32 des kantonalen Gesetzes vom 23. Mai 1989 Urteil des Verwaltungsgerichts des Kantons Bern vom 6. Dez. 2022, UV/21/844, Seite 4 über die Verwaltungsverfahrensregeln [VRPG; BSG 155.21]) eingehalten sind, ist auf die Beschwerde einzutreten. 1.2 Anfechtungsobjekt bildet der Einspracheentscheid vom 5. November 2021 (act. II 45). Streitig und zu prüfen ist, ob die Beschwerdegegnerin ihre Leistungen im Zusammenhang mit dem Ereignis vom 19. Juni 2020 zu Recht per 30. September 2020 eingestellt hat. 1.3 Die Abteilungen urteilen gewöhnlich in einer Kammer bestehend aus drei Richterinnen oder Richtern (Art. 56 Abs. 1 GSOG). 1.4 Das Gericht überprüft den angefochtenen Entscheid frei und ist an die Begehren der Parteien nicht gebunden (Art. 61 lit. c und d ATSG; Art. 80 lit. c Ziff. 1 und Art. 84 Abs. 3 VRPG). 2. 2.1 Die Zusprechung von Leistungen der obligatorischen Unfallversicherung setzt grundsätzlich das Vorliegen eines Berufsunfalles, eines Nichtberufsunfalles oder einer Berufskrankheit (Art. 6 Abs
Predicted	considerations: Erwägungen: 1. 1.1 Der angefochtene Entscheid ist in Anwendung von Sozialversicherungsrecht ergangen. Die Sozialversicherungsrechtliche Abteilung des Verwaltungsgerichts beurteilt gemäss Art. 57 des Bundesgesetzes über den Allgemeinen Teil des Sozialversicherungsrechts vom 6. Oktober 2000 (ATSG; SR 830.1) i.V.m. Art. 54 Abs. 1 lit. a des kantonalen Gesetzes über die Organisation der Gerichtsbehörden und der Staatsanwaltschaft vom 11. Juni 2009 (GSOG; BSG 161.1) Beschwerden gegen solche Entscheide. Die Beschwerdeführerin ist im vorinstanzlichen Verfahren mit ihren Anträgen nicht durchgedrungen, durch den angefochtenen Entscheid berührt und hat ein schutzwürdiges Interesse an dessen Aufhebung, weshalb sie zur Beschwerde befugt ist (Art. 59 ATSG). Die örtliche Zuständigkeit ist gegeben (Art. 58 ATSG). Da auch die Bestimmungen über Frist (Art. 60 ATSG) sowie Form (Art. 61 lit. b ATSG; Art. 81 Abs. 1 i.V.m. Art. 32 des kantonalen Gesetzes über die Verwaltungsverfahrensregeln vom 23. Mai 1989 [VRPG; BSG 155.21]) eingehalten sind, ist auf die Beschwerde einzutreten. 1.2 Das Gericht überprüft den angefochtenen Entscheid frei und ist an die Begehren der Parteien nicht gebunden (Art. 61 lit. c und d ATSG; Art. 80 lit. c Ziff. 1 und Art. 84 Abs. 3 VRPG). 1.3 Der Beschwerdeführer hat am vorinstanzlichen Verfahren teilgenommen, ist durch den angefochtenen Entscheid besonders berührt und hat ein schutzwürdiges Interesse an dessen Aufhebung oder Änderung. Auf die Beschwerde ist einzutreten. 2. 2.1 Die Beschwerdegegnerin hat die Bestimmungen auf den Anspruch auf Heilbehandlung und Taggeld in Form von Heilbehandlung zufolge ihrer im August 2021 (act. II 42 und 46) eingereichten medizinischen Berichte zutreffend wiedergegeben. Darauf wird verwiesen. 2.2 Nach der Rechtsprechung ist der Beschwerdeführer durch den angefochtenen Entscheid berührt und haben ein schutzwürdiges
Scores	BERT: 91.69, BLEU: 84.17, METEOR: 62.43, ROUGE-1/2/L: 75.49 / 65.78 / 75.49
Target	considerations: EN DROIT I. Conformément à l'art. 134 al. 1 let. a ch. 2 de la loi sur l'organisation judiciaire, du 26 septembre 2010 (LOJ - E 2 05 ) en vigueur dès le 1 er janvier 2011, la chambre des assurances sociales de la Cour de justice connaît, en instance unique, des contestations prévues à l'art. 56 de la loi fédérale sur la partie générale du droit des assurances sociales, du 6 octobre 2000 (LPGA - RS 830.1 ) relatives à la loi fédérale sur l'assurance-invalidité du 19 juin 1959 (LAI - RS 831.20 ).!.[endif]>![if] Sa compétence pour juger du cas d'espèce est ainsi établie. 2. A teneur de l'art. 1 al. 1 LAI, les dispositions de la LPGA s'appliquent à l'assurance-invalidité (ci-après : AI), à moins que la loi n'y déroge expressément.!.[endif]>![if] Toutefois, les modifications légales contenues dans la LPGA constituent, en règle générale, une version formalisée dans la loi de la jurisprudence relative aux notions correspondantes avant l'entrée en vigueur de la LPGA ; il n'en découle aucune modification du point de vue de leur contenu, de sorte que la jurisprudence développée à leur propos peut être reprise et appliquée (ATF 130 V 343 consid. 3). 3. Les modifications de la loi fédérale sur l'assurance-invalidité du 21 mars 2003 (4 <sup>ème</sup> révision), du 6 octobre 2006 (5 <sup>ème</sup> révision) et du 18 mars 2011 (révision 6a), entrées en vigueur le 1 er janvier 2004, respectivement, le 1 er janvier 2008 et le 1 er janvier 2012, entraînent la modification de nombreuses dispositions légales dans le domaine de l'assurance-invalidité.
Predicted	considerations: EN DROIT I. Conformément à l'art. 134 al. 1 let. a ch. 4 de la loi sur l'organisation judiciaire, du 26 septembre 2010 (LOJ; RS E 2 05 ) en vigueur dès le 1 er janvier 2011, la Chambre des assurances sociales de la Cour de justice connaît, en instance unique, des contestations prévues à l'art. 56 de la loi fédérale sur la partie générale du droit des assurances sociales, du 6 octobre 2000 (LPGA; RS 830.1 ) relatives à la loi fédérale sur l'assurance-invalidité du 19 juin 1959 (LAI; RS 831.20 ).!.[endif]>![if] Sa compétence pour juger du cas d'espèce est ainsi établie.!.[endif]>![if] 2. Les dispositions de la LPGA s'appliquent à l'assurance-invalidité, à moins que la loi n'y déroge expressément.!.[endif]>![if] Toutefois, les modifications légales contenues dans la LPGA constituent, en règle générale, une version formalisée dans la loi de la jurisprudence relative aux notions correspondantes avant l'entrée en vigueur de la LPGA ; il n'en découle aucune modification du point de vue de leur contenu, de sorte que la jurisprudence développée à leur propos peut être reprise et appliquée (ATF 130 V 230 consid. 1.1; 335 consid. 1.2; ATF 129 V 4 consid. 1.2; ATF 129 V 4 consid. 1.2; ATF 127 V 467 consid. 1, 126 V 136 consid. 4b et les références). 3. A teneur de l'art. 17 al. 1 de la loi fédérale sur la partie générale du droit des assurances sociales, du 6 octobre 2000 (LPGA; RS 830.1 ), les modifications légales contenues dans la LPGA constituent, en règle générale, une version formalisée dans la LAI.
Scores	BERT: 91.27, BLEU: 85.8, METEOR: 68.12, ROUGE-1/2/L: 78.29 / 66.82 / 78.29
Target	considerations: in diritto In ordine 2.1. La presente vertenza non pone questioni giuridiche di principio e non è di rilevante importanza (ad esempio per la difficoltà dell'istruttoria o della valutazione delle prove). Il TCA può dunque decidere nella composizione di un Giudice unico ai sensi dell'articolo 49 cpv. 2 LOG (STF 9C_699/2014 del 31 agosto 2015; 8C_855/2010 dell'11 luglio 2011; 9C_211/2010 del 18 febbraio 2011). Nel merito 2.2. Oggetto del contendere è sapere se a ragione l'Ufficio AI ha soppresso all'assicurato il diritto ad una rendita intera con effetto dal 1. settembre 2015 (cfr. consid. 1.3). L'insorgente postula l'annullamento della decisione impugnata del 10 maggio 2017 e quindi il mantenimento del diritto alla rendita intera. 2.3. Secondo l'art. 4 cpv. 1 LAI in relazione con gli art. 7 e 8 della LPGA, con invalidità s'intende l'incapacità al guadagno presunta permanente o di rilevante durata, cagionata da un danno alla salute fisica o psichica, conseguente ad infermità congenita, malattia o infortunio. Gli elementi fondamentali dell'invalidità, secondo la surriferita definizione, sono quindi un danno alla salute fisica o psichica conseguente a infermità congenita, malattia o infortunio, e la conseguente incapacità di guadagno. Occorre quindi che il danno alla salute abbia cagionato una diminuzione della capacità di guadagno, perché il caso possa essere sottoposto all'assicurazione per l'invalidità ( Duc, L'assurance invalidité, in: Meyer (Hrsg.), Schweizerisches Bundesverwaltungsrecht, Band XIV, Soziale Sicherheit, 2a ed., Basilea/Ginevra/Monaco di Baviera 2007, pag. 1411, n. 46 ). Giusta l'art. 28 cpv.
Predicted	considerations: in diritto In ordine 2.1. La presente vertenza non pone questioni giuridiche di principio e non è di rilevante importanza (ad esempio per la difficoltà dell'istruttoria o della valutazione delle prove). Il TCA può dunque decidere nella composizione di un Giudice unico ai sensi degli articoli 26 e cpv. 2 della Legge organica giudiziaria civile e penale e 2 cpv. 1 della Legge di procedura per le cause davanti al Tribunale delle assicurazioni (cfr. STF 9C_211/2010 del 18 febbraio 2011; STF 9C_792/2007 del 7 novembre 2008; STF H 180/06 e H 183/06 del 21 dicembre 2007; STFA I 707/00 del 21 luglio 2003; STFA H 335/00 del 18 febbraio 2002; STFA U 347/98 del 10 ottobre 2001, pubblicata in RDAT I-2002 pag. 190 seg.; STFA H 304/99 del 22 dicembre 2000; STFA I 623/98 del 26 ottobre 1999). Nel merito 2.2. Oggetto della lite è sapere se l'Ufficio AI ha diritto alla rendita intera intera dal 1. settembre 2015 (doc. AI 122/445-450), a causa dell'invalidità della persona assicurata. Secondo l'art. 4 cpv. 1 LAI in relazione con gli art. 7 e 8 della LPGA, con invalidità s'intende l'incapacità al guadagno presunta permanente o di rilevante durata, cagionata da un danno alla salute fisica o psichica, conseguente a infermità congenita, malattia o infortunio. Gli elementi fondamentali dell'invalidità, secondo la surriferita definizione, sono quindi un danno alla salute fisica o psichica conseguente a infermità congenita, malattia o infortunio, e la conseguente incapacità di guadagno. Occorre quindi che il danno alla salute abbia cagionato una diminuzione della capacità di guadagno, perché il caso possa
Scores	BERT: 86.57, BLEU: 81.39, METEOR: 65.76, ROUGE-1/2/L: 68.54 / 58.87 / 66.85

Target	considerations: Das Bundesverwaltungsgericht zieht in Erwägung: 1. Nach Art. 53 Abs. 1 KVG kann gegen Beschlüsse der Kantonsregierungen nach Art. 39 KVG beim Bundesverwaltungsgericht Beschwerde geführt werden. Der angefochtene Beschluss Nr. 617/2019 des Regierungsrats des Kantons Zürich vom 26. Juni 2019 wurde gestützt auf Art. 39 KVG erlassen. Das Bundesverwaltungsgericht ist deshalb zur Beurteilung der Beschwerde zuständig (vgl. auch Art. 90a Abs. 2 KVG). Die Beschwerdeführerin hat am vorinstanzlichen Verfahren teilgenommen, ist als Adressatin durch den angefochtenen Regierungsratsbeschluss besonders berührt und hat insoweit an dessen Aufhebung beziehungsweise Abänderung ein schutzwürdiges Interesse (Art. 48 Abs. 1 VwVG). Sie ist daher zur Beschwerde legitimiert. Auf die frist- und formgerecht erhobene Beschwerde ist, nachdem auch der Kostenvorschuss rechtzeitig geleistet wurde, einzutreten (Art. 50 Abs. 1, Art. 52 Abs. 1 und Art. 63 Abs. 4 VwVG). Das Hauptrechtsbegehren der Beschwerdeführerin richtet sich auf Aufhebung der Ziffer I des Dispositivs des angefochtenen Beschlusses. Soweit sie dabei Bezug auf «C. Neue Regelung der Verlegungen ab Geburtshaus» nimmt, ist das im Lichte der Begründung der Beschwerde nicht so zu verstehen, dass sie auch die Aufhebung eines Teils der Begründung des angefochtenen Beschlusses beantragt. Der Antrag der Vorinstanz, auf die Beschwerde sei teilweise nicht einzutreten, weil sie sich auch gegen die Begründung des angefochtenen RRB richtet, ist daher unbegründet. 2. 2.1 Mit Beschwerde gegen einen Spitalistenentscheid im Sinne von Art. 39 Abs. 1 KVG kann die Verletzung von Bundesrecht unter Einschluss des Missbrauchs oder der Überschreitung des Ermessens und die unrichtige oder unvollständige Feststellung des rechtserheblichen Sachverhalts gerügt werden. Nicht zulässig ist hingegen die Rüge der
Predicted	considerations: Das Bundesverwaltungsgericht zieht in Erwägung: 1. 1.1 Gemäss Art. 31 des Verwaltungsgerichtsgesetzes vom 17. Juni 2005 (VGG, SR 173.32) beurteilt das Bundesverwaltungsgericht Beschwerden gegen Verfügungen nach Art. 5 des Verwaltungsverfahrensgesetzes vom 20. Dezember 1968 (VwVG, SR 172.021), sofern keine Ausnahme nach Art. 32 VGG vorliegt. Als Vorinstanzen gelten die in Art. 33 und 34 VGG genannten Behörden. Dazu gehören u.a. Verfügungen im Sinne von Art. 5 VwVG. Das Bundesverwaltungsgericht ist somit zuständig für die Beurteilung der vorliegenden Beschwerde und entscheidet auf dem Gebiet des Gesundheitsschutzes endgültig, ausser bei Eintreten auf einen Entscheid, der eine Verfügung im Sinne von Art. 5 VwVG betrifft. Eine Ausnahme nach Art. 32 VGG liegt nicht vor. 1.2 Das Bundesverwaltungsgericht ist zur Beurteilung der vorliegenden Beschwerde zuständig. 1.3. Gegen Entscheide des Regierungsrats ist ein Beschwerdeverfahren zulässig, wenn dieses eine Rechtsmittelinstanz ist und grundsätzlich der Rechtsschutzinteresse der Beschwerdeberechtigten gegeben ist (vgl. Art. 32 VGG; BGE 138 II 274 E. 5.1 S. 280). 1.4 Das Bundesverwaltungsgericht wendet im Beschwerdeverfahren das Bundesrecht von Amtes wegen an (Art. 106 Abs. 1 des Bundesgesetzes vom 20. Dezember 1968 über das Verwaltungsverfahren [VwVG, SR 172.021]). Es wird gemäss Art. 62 Abs. 4 VwVG zur Beurteilung von Beschwerden in Zivilsachen gemäss Art. 82 Abs. 1 und Art. 86 Abs. 1 Bst. d bis Bst. e VwVG grundsätzlich nur geprüft, wenn diese vor Bundesgericht nicht mehr vorgetragen wurden und keine Verletzung des Opfers gegeben wurde. Das Bundesverwaltungsgericht wendet im Beschwerdeverfahren das Bundesrecht von Amtes wegen an. Es ist gemäss Art. 62 Abs. 4 VwVG an die Begründung der Begehren nicht gebunden und kann die Beschwerde auch aus anderen als den geltend gemachten Gründen gutheissen
Scores	BERT: 75.72, BLEU: 67.17, METEOR: 31.26, ROUGE-1/2/L: 26.43 / 8.84 / 25.23
Target	considerations: Considérant en droit: 1. Selon l'art. 5 al. 1 des dispositions du RPAC relatives à la Commission de recours individuel, la décision rendue par cette autorité peut faire l'objet d'un recours de droit administratif auprès du Tribunal cantonal dans les trente jours suivant la communication de la décision motivée, conformément à la loi cantonale du 28 octobre 2008 sur la procédure administrative (LPA-VD; BLV 173.36). D'après l'art. 92 al. 1 LPA-VD, le Tribunal cantonal connaît des recours contre les décisions et décisions sur recours rendues par les autorités administratives, lorsque la loi ne prévoit aucune autre autorité pour en connaître. En l'espèce, le recours a été déposé en temps utile (art. 5 al. 1 RPAC) et il satisfait aux autres conditions formelles de recevabilité (art. 75 al. 1 let. a, 79 al. 1 et 99 LPA-VD, applicables par renvoi de l'art. 5 al. 1 RPAC), si bien qu'il convient d'entrer en matière. 2. a) La classification salariale du poste qu'occupe le recourant est litigieuse en l'occurrence. b) L'organisation de l'administration fait partie des tâches propres des autorités communales (art. 2 de la loi vaudoise du 28 février 1956 sur les communes [LC; BLV 175.11]). Selon cette loi, il incombe au Conseil général ou communal de définir le statut des collaborateurs communaux et la base de leur rémunération (art. 4 al. 1 ch. 9 LC), la municipalité ayant la compétence de nommer les collaborateurs et employés de la commune, de fixer leur traitement et d'exercer le pouvoir disciplinaire (art. 42 al. 1 ch. 3 LC). Conformément à l'art. 33 al. 1 RPAC, le traitement du fonctionnaire comprend le traitement de base (let. a), les allocations
Predicted	considerations: Considérant en droit: 1. 1.1 Le personnel doit maintenir, par contrat de droit privé, une commune, un partenaire public, pour assurer la bonne exécution de leurs missions spéciales, ainsi qu'un minimum d'administration (art. 10 al. 1 du RPAC). L'art. 10 al. 2 RPAC prévoit que les fonctionnaires dans toute la ville de Lausanne sont soumis à la qualification communale et leur remplacement a lieu par des fonctions communales cantonales, comme applicable en la matière. Elle constitue en revanche une décision sujette à recours auprès de la juridiction compétente (art. 20 al. 2 et art. 30 al. 1 de la loi sur le personnel des communes du 15 novembre 1976 (LPAC). 1.2 Selon la jurisprudence relative aux dispositions transitoires applicables au moment de la notification de la décision pour connaître du jugement du service, la décision de rémunération fait apparaître un changement de système de rémunération selon le RPAC. Cette disposition, qui comprend ainsi des critères objectifs et suffisants, permet au fonctionnaire de connaître de tous les aspects de la classification (Tappy, CPC, n. 1 ad art. 10 LPAC), ainsi que par le service de réduire son salaire. En l'espèce, l'autorité intimée a constaté que l'établissement des faits importants du dossier et de la satisfaction exigée sur le plan d'administration devait être complété dans la mesure où cela n'était pas possible, ce qui constituait effectivement un changement de système (Tappy, op. cit., n. 13 ad art. 10 LPAC). Il a constaté que le Service du personnel avait fixé le niveau de fonction attribué au fonctionnaire pour son titre de travail. Il a en outre estimé que cette classification était une décision de mise en œuvre déléguée à celui qui a entrepris une modification légale ou réglementaire du service. Parcon
Scores	BERT: 75.74, BLEU: 66.19, METEOR: 31.69, ROUGE-1/2/L: 28.32 / 4.21 / 26.01
Target	considerations: Diritto: 1. Il ricorso in materia di diritto pubblico può essere presentato per violazione del diritto, così come stabilito dagli art. 95 e 96 LTF. Il Tribunale federale fonda la sua sentenza sui fatti accertati dall'autorità inferiore (art. 105 cpv. 1 LTF). Può scostarsi da questo accertamento solo qualora esso sia avvenuto in modo manifestamente inesatto, ovvero arbitrario (DTF 134 V 53 consid. 4.3 pag. 62), oppure in violazione del diritto ai sensi dell'art. 95 LTF (art. 105 cpv. 2 LTF). Salvo i casi in cui tale inesattezza sia lampante, la parte ricorrente che intende contestare i fatti accertati dall'autorità inferiore deve spiegare, in maniera circostanziata, per quale motivo ritiene che le condizioni di una delle eccezioni previste dall'art. 105 cpv. 2 LTF sarebbero realizzate (cfr. DTF 133 II 249 consid. 1.4.3 pag. 254 con riferimento). 2. Oggetto del contendere è sapere se sia lesivo del diritto federale il giudizio cantonale, che conferma la decisione su opposizione, la quale a sua volta nega all'assicurato le indennità per insolvenza per non avere rivendicato tempestivamente le proprie pretese salariali. 3. 3.1. Il Tribunale cantonale delle assicurazioni, esposto lo svolgimento del processo, le norme e la prassi ritenute applicabili, ha accertato che il ricorrente è stato legato da un contratto di lavoro con la ditta B. SA, dichiarata fallita il 15 maggio 2018, dal 15 aprile 2017 al 18 settembre 2017 come informatico e con un salario di fr. 3'360.- mensili. L'ultimo giorno di lavoro effettuato è stato il 1° settembre 2017. Il salario è stato saldato dal 15 al 30 aprile 2017. Per contro, non è stato pagato per i mesi di maggio, giugno, luglio e agosto, nonché per il 1° settembre 2017. Il
Predicted	considerations: Diritto: 1. 1.1. La procedura ricorsuale in materia di diritto pubblico è disciplinata dal diritto federale (art. 82 lett. a LTF) e dalla legge federale, in vigore dal 1° gennaio 2003, sull'assicurazione contro la disoccupazione: si tratta infatti di una causa di diritto pubblico contro una decisione finale adottata dalla Corte cantonale delle assicurazioni, entrata in vigore il 1° gennaio 2003. Il ricorso presentato dall'assoggettato al diritto di ricorrere (art. 89 cpv. 1 LTF) è di massima ammissibile dinanzi al Tribunale federale (art. 89 cpv. 1 LTF). 1.2. Il ricorso in materia di diritto pubblico può essere presentato per violazione del diritto, così come stabilito dagli art. 95 e 96 LTF, oltre che per violazione del diritto ai sensi dell'art. 95 LTF. Il Tribunale federale applica il diritto federale (art. 106 cpv. 1 LTF), senza essere vincolato né dai motivi addotti nel ricorso (art. 106 cpv. 2 LTF). Per contro, nel ricorso in materia di diritto pubblico il Tribunale federale esamina d'ufficio e con piena cognizione l'ammissibilità dei gravami che gli vengono sottoposti (DTF 133 III 439 consid. 1.3). 1.3. La critica del giudizio impugnato esplica degli effetti (art. 105 cpv. 1 LTF). Non è ammissibile che i ricorsi in materia di diritto pubblico possano essere decisi in base al diritto federale, ai sensi dell'art. 95 LTF, senza istruttoria (art. 97 cpv. 1 LTF; art. 105 cpv. 2 LTF). 1.3. Il Tribunale federale esamina d'ufficio e con piena cognizione l'ammissibilità dei ricorsi che gli vengono sottoposti (DTF 133 III 439 consid. 1.3). 1.4. Con il ricorso in materia di diritto pubblico contro la decisione di primo grado, il Tribunale cantonale ha emesso
Scores	BERT: 75.79, BLEU: 66.28, METEOR: 30.29, ROUGE-1/2/L: 37.74 / 20.43 / 36.48

Target	considerations: Das Versicherungsgericht zieht in Erwägung: 1. Streitig und zu prüfen ist der Rentenanspruch der Beschwerdeführerin. - 3 - 2. Am 1. Januar 2022 sind die Änderungen betreffend Weiterentwicklung der IV (WEIV) in Kraft getreten. Weder dem IVG noch der IVV sind besondere Übergangsbestimmungen betreffend die Anwendbarkeit dieser Änderungen im Hinblick auf nach dem 1. Januar 2022 beurteilte mögliche Ansprüche des Zeitraums bis zum 31. Dezember 2021 zu entnehmen. Es sind daher nach den allgemeinen übergangsrechtlichen Grundsätzen jene Bestimmungen anzuwenden, die bei der Erfüllung des rechtlich zu ordnenden oder zu Rechtsfolgen führenden Tatbestands Geltung haben beziehungsweise hatten (vgl. Urteil des Bundesgerichts 8C_136/2021 vom 7. April 2022 E. 3.2.1 mit Hinweis unter anderem auf BGE 144 V 210 E. 4.3.1 S. 213). Da vorliegend Leistungen mit Anspruchsbeginn vor dem 1. Januar 2022 streitig sind, ist für deren Beurteilung die bis zum 31. Dezember 2021 geltende Rechtslage massgebend. 3. 3.1. Die Beschwerdegegnerin stützte sich in der Verfügung vom 14. Juni 2022 (Vernehmlassungsbeilage [VB] 162) in medizinischer Hinsicht im Wesentlichen auf das bidisziplinäre Gutachten der SMAB vom 3. August 2021 (VB 139; Fachdisziplinen: Neurologie und Psychiatrie). 3.2. Die SMAB-Gutachterinnen stellten in der interdisziplinären Gesamtbeurteilung die folgenden Diagnosen mit Auswirkungen auf die Arbeitsfähigkeit (VB 139.1 S. 6): "1. Neuropathische Schmerzen an der Innenseite des Mittelfingers links und der Handinnenfläche links (Schnittverletzung am Mittelfinger links mit initialer Nervenverletzung 19.11.2014, epineurale und weitere mehrfache operative Eingriffe 2014-2018), 2. Chronische Schmerzstörung mit somatischen und psychischen (ICD-10: F45.41)". Ohne Auswirkung auf die Arbeitsfähigkeit sei eine rezidivierende depressive Störung, gegenwärtig remittiert. Die Gutachterinnen hielten fest, insgesamt bestehe aufgrund der psychiatrischen und neurologischen
Predicted	considerations: Das Gericht zieht in Erwägung: 1. Gemäss Art. 87 Abs. 2 des Bundesgesetzes über die Unfallversicherung (UVG) besteht Anspruch auf eine ganze Rente, wenn die versicherte Person mindestens zu 66 2/3 Prozent, auf eine halbe Rente, wenn sie mindestens zu 50 Prozent oder auf eine Viertelsrente, wenn sie mindestens zu 40 Prozent invalid ist. In Härtefällen besteht gemäss Art. 87 Abs. 2 UVG auch ein Anspruch auf eine halbe Rente, wenn sie mindestens zu 50 Prozent, wenn sie mindestens zu 40 Prozent invalid ist (Art. 8 Abs. 1 UVG). Anspruch auf eine Rente haben gemäss Art. 28 Abs. 2 des Bundesgesetzes über die Invalidenversicherung (IVG) Versicherte, die im Sinne von Art. 16 Abs. 1 IVG in Verbindung mit Art. 28 bis Abs. 1 IVG in Verbindung mit Abs. 1 und 2 IVG auch dann Anspruch auf eine Viertelsrente, wenn sie mindestens zu 40 Prozent invalid sind (Art. 28 Abs. 2 IVG i.V.m. Art. 16 ATSG). 2. 2.1 Anspruch auf eine Rente haben gemäss Art. 28 Abs. 1 IVG Versicherte, die: a. ihre Erwerbsfähigkeit oder die Fähigkeit, sich im Aufgabenbereich zu betätigen, nicht durch zumutbare Eingliederungsmassnahmen wieder herstellen, erhalten oder verbessern können; b. während eines Jahres ohne wesentlichen Unterbruch durchschnittlich mindestens 40 Prozent arbeitsunfähig (Art. 6 ATSG) gewesen sind; und c. nach Ablauf dieses Jahres zu mindestens 40 Prozent invalid (Art. 8 ATSG) sind. 2.2 Die Beschwerdeführerin bringt vor, die ärztlich eingeholten ärztlichen Berichte seien als diagnostisch zu qualifizieren. Das trifft vorliegend nicht zu. Ihr Gesundheitszustand sei gemäss Abklärungen vom RAD mit einer Invalidität von mindestens 40 Prozent zu vereinbaren. Die Leistungsfähigkeit sei in Art. 16 ATSG eingetreten. 2.2 Die Leistungsfähigkeit sei in Art. 16 ATSG i.V.m. Art. 28 Abs. 1 IVG
Scores	BERT: 62.42, BLEU: 51.67, METEOR: 19.69, ROUGE-1/2/L: 17.58 / 2.56 / 17.58
Target	considerations: le conseil de B, a rappelé que celle-ci contestait que son frère aurait été en droit de loger gratuitement au 1, dès le mois d'avril 2015. Afin d'éviter la notification d'une poursuite interruptive de la prescription, A, était invité à signer une déclaration de renonciation à la prescription concernant les loyers qui, selon B, étaient dus à la succession de leurs défunts parents. Etait annexée à ce courrier une déclaration de renonciation à invoquer la prescription libellée en ces termes : " Par la présente, A, déclare renoncer à invoquer toute exception de prescription vis-à-vis des successions de [feu D, et feu C,], eu égard à toutes prétentions éventuelles de loyers et charges alors qu'il logeait chemin 1, à Fribourg. Cette déclaration de renonciation est valable jusqu'au 31 décembre 2020. La présente déclaration ne vaut en aucun cas reconnaissance de dette ou de responsabilité quelconque. La présente déclaration n'a d'effet que si la prescription n'est pas déjà acquise à ce jour ". A, n'a pas donné suite à ce courrier. f. Par réquisition de poursuite du 7 avril 2020, reçue le lendemain par l'Office cantonal des poursuites de Genève (ci-après : l'Office), Me H, agissant en qualité de représentant du créancier, à savoir " Pour l'hoirie de feu C, décédée le 2019, Madame B, Rte 2, Fribourg ", a requis la poursuite de A, domicilié avenue 3, Genève, pour la somme de 112'500 fr. avec intérêts à 5% dès le 1er mai 2017, réclamée à titre de " Loyers et charges dus pour le logement du 1, à Fribourg d'avril 2015 à mai 2019 (50 x
Predicted	considerations: EN DROIT I. 1.1 Le litige porte sur la désignation d'un représentant de la communauté héréditaire et la désignation d'un administrateur de la régie G, SA (art. 602 al. 3 CC). Le motif de la désignation du représentant est incontestable. Selon l'art. 602 al. 5 CC, la désignation d'un représentant de la communauté héréditaire est donc une décision de dernière instance cantonale au sens des art. 80 al. 1 et 82 al. 1 CPC, qui doit être notifiée à l'unique partie plaignante qui a pris part à la procédure (ATF 110 IV 92 consid. 1a; arrêt du Tribunal fédéral X, du 27 juillet 2004 ; ATF 99 III 58 consid. 1; ATF 118 Ia 397 consid. 1b; Tappy, CPC-VD, n. 3 ad art. 602 CC ; Replin, Le représentant de la communauté héréditaire, 5 <sup>ème</sup> éd., Lausanne 2013, p. 569; ATF 118 IV 286 consid. 2a; TF 6B_211/2007 du 29 janvier 2008, consid. 5.3; ATF 117 IV 29 consid. 3b; TF 8B_44/2007 du 15 août 2007, consid. 3.2; Tappy, Procédure civile, tome II, ad art. 602; Piquerez, in : Kuhn/Jeanneret [éd.], Basler Kommentar, n. 6 ad art. 602; TF 7B_51/2007 du 1 <sup>er</sup> janvier 2008, consid. 3.2; ATF 130 III 136 consid. 1.2.1; TF 9C_438/2007 du 30 septembre 2007, consid. 3.1; ATF 134 III 102 consid. 3.1; TF 6B_71/2007 du 24 août 2007, consid. 5b; TF 9C_792/2007 du 11 août 2007, consid. 4.2). 1.2 La désignation d'un représentant et de l'administration régulière de l'ensemble de la succession ont été rejetées. 2. 2.1. L'art. 602 al. 3 CC ouvre un recours au Tribunal cantonal, sans être lié par l'
Scores	BERT: 63.90, BLEU: 46.65, METEOR: 19.12, ROUGE-1/2/L: 11.52 / 1.64 / 10.91
Target	considerations: A.c. Con progetto di decisione dell'8 gennaio 2014 l'UAI ha concluso, in via di riconsiderazione, per la soppressione della mezza rendita d'invalidità con effetto retroattivo al 30 aprile 2011. In sede di audizione, l'UAI ha esperito nuovi accertamenti, in particolare la perizia pluridisciplinare dell'8 aprile 2016. Con decisione del 14 luglio 2016 l'UAI ha riconosciuto che al momento dell'assegnazione della mezza rendita d'invalidità è incorso in un errore manifesto avendo applicato il "Prozentvergleich" in luogo del metodo ordinario di confronto dei redditi, da cui risulta un grado d'invalidità del 26 %. Ritenuta l'importanza della sua correzione, l'UAI, riconsiderando le decisioni 14 dicembre 2006 e 14 maggio 2007 nonché la comunicazione del 20 marzo 2009, ha predisposto la soppressione del diritto alla mezza rendita invalidità con effetto retroattivo al 30 aprile 2011. B. A, si è aggravato il 13 settembre 2016 al Tribunale delle assicurazioni del Cantone Ticino, che con giudizio del 17 maggio 2017 ha respinto il gravame. C. Il 19 giugno 2017 (timbro postale) A, inoltra un ricorso in materia di diritto pubblico al Tribunale federale, cui chiede di ammettere il ricorso e riformare il giudizio cantonale nel senso di confermare il diritto ad almeno una mezza rendita d'invalidità e ordinare all'UAI di versare tutte le prestazioni trattenute dal momento della revisione ad oggi. L'UAI e l'Ufficio federale delle assicurazioni sociali rinunciano a determinarsi. Diritto: 1. Il ricorso in materia di diritto pubblico può essere presentato per violazione del diritto così come stabilito dagli art. 95 e 96 LTF. Il Tribunale federale esamina d'ufficio l'applicazione del diritto federale (art. 106 cpv. 1 L
Predicted	considerations: Diritto: 1. Il Tribunale delle assicurazioni esamina d'ufficio e con piena cognizione l'ammissibilità dei gravami che gli vengono sottoposti (DTF 134 V 131 consid. 1.1 pag. 131; 133 V 477 consid. 1.1 pag. 481). In particolare, i gravami inoltrati già in tempo utile contro le decisioni di prima istanza rese dall'UAI in materia di assicurazioni sociali possono essere decisi sulla scorta degli atti, senza istruttoria (DTF 133 V 477 consid. 1.2 pag. 481). Il giudizio può essere reso sulla base degli atti, senza istruttoria (DTF 133 V 477 consid. 1.2 pag. 481; 133 V 477 consid. 1.2 pag. 481). 2. Il Tribunale delle assicurazioni esamina d'ufficio e con piena cognizione l'ammissibilità dei gravami che gli vengono sottoposti (DTF 133 V 477 consid. 1.2 pag. 481). 3. 3.1. Giusta l'art. 4 cpv. 1 LAI in relazione con gli art. 7 e 8 della LPGA, con invalidità s'intende l'incapacità al guadagno presunta permanente o di rilevante durata, cagionata da un danno alla salute fisica o psichica, conseguente a infermità congenita, malattia o infortunio. Gli elementi fondamentali dell'invalidità, secondo la surriferita definizione, sono quindi un danno alla salute fisica o psichica conseguente a infermità congenita, malattia o infortunio, e la conseguente incapacità di guadagno. Occorre quindi che il danno alla salute abbia cagionato una diminuzione della capacità di guadagno, perché il caso possa essere sottoposto all'assicurazione per l'invalidità (DTF 134 V 329 consid. 3.2.3 pag. 331; 133 V 477 consid. 3.3.1 pag. 481; 129 V 4 consid. 1.2 pag. 5 con rinvii). 3.2. Giusta
Scores	BERT: 67.86, BLEU: 55.76, METEOR: 25.41, ROUGE-1/2/L: 24.32 / 5.59 / 23.65

Table 15: Three examples of generated summaries by mTS<sub>Base</sub> with input length 4096, showcasing high, average, and low scored outputs in LDS Task

Target	regeste: Regeste Teilklage und negative Feststellungswiderklage ( Art. 86 und 224 ZPO ). Die Ausnahme vom Erfordernis der gleichen Verfahrensort gemäss Art. 224 Abs. 1 ZPO für negative Feststellungswiderklagen ist nicht auf den Fall beschränkt, dass es sich bei der Hauptklage um eine sogenannte echte Teilklage handelt, sondern gilt allgemein dann, wenn die Teilklage eine Ungewissheit zur Folge hat, die es rechtfertigt, die Feststellung des Nichtbestands einer Forderung oder eines Rechtsverhältnisses zu verlangen (E. 2).
--------	---

Predicted	regeste: Regeste Voraussetzungen, unter denen eine Widerklage erhoben werden darf, wenn eine echte Teilklage beurteilt wird (E. 2.1). Zur Anwendung von Art. 224 Abs. 1 ZPO ist die Ausnahme vom Erfordernis der gleichen Verfahrensart gemäss Art. 224 Abs. 1 ZPO nicht auf diesen Fall beschränkt, sondern gilt allgemein dann, wenn die Teilklage eine Ungewissheit zur Folge hat, die es rechtfertigt, im Sinne von Art. 88 ZPO die Feststellung des Nichtbestands einer Forderung oder eines Rechtsverhältnisses zu verlangen (E. 2.2).
Scores	BERT: 88.03, BLEU: 30.45, METEOR: 43.78, ROUGE-1/2/L: 72.44 / 54.05 / 70.87
Target	regeste: Regeste Uferschutzplanung Wohlensee; Gesamtinteressenabwägung. Die Behörden nehmen bei der Genehmigung einer Sondernutzungsplanung eine umfassende Interessenabwägung vor (vgl. Art. 3 RPV ; E. 3.2). Würdigung des Gutachtens der Vogelwarte Sempach und der Stellungnahme des BAFU als Fachbehörde (E. 5.5, 6.3 und 6.4). Im zu beurteilenden Fall wird das öffentliche Interesse an einer ufernahen Wegführung (vgl. Art. 3 Abs. 2 lit. c RPG) relativiert, weil ein durchgehender Sichtschutz zum Schutz der im Gebiet vorkommenden Vogelarten erforderlich wäre, was Spaziergängern das Erleben der Uferlandschaft nur sehr beschränkt ermöglichen würde (E. 6.5 sowie 3.3). Das Gebiet des Wohlensees stellt eines von insgesamt 25 Wasser- und Zugvogelreservaten von nationaler Bedeutung dar (vgl. Art. 11 Abs. 2 JSG, Art. 2 Abs. 1 i.V.m. Anhang 1 WZVV).
Predicted	regeste: Regeste Art. 3 Abs. 2 lit. c RPG ; schweizerische Vogelwarte Sempach; Umweltschutz. Gutachten im Bereich des Umweltrechts, insbesondere von dem UVG/USG ergangenen Streitwert (E. 3.1). Ein öffentlicher Weg, der über Privatgrundstücke führt und mit dem Enteignungsrecht zugunsten des Gemeinwesens verbunden ist, stellt eine öffentlich-rechtliche Eigentumsbeschränkung dar (E. 3.3). Interessenabwägungen sind grundsätzlich frei (E. 3.4 und 3.5). Umfang des geplanten Uferweges; ein ausweisenes öffentliches Interesse an einem direkt dem Ufer entlangführenden Weg sowie an einer zumindest ufernahen Wegführung (E. 3.5).
Scores	BERT: 73.91, BLEU: 47.21, METEOR: 30.91, ROUGE-1/2/L: 32.75 / 10.63 / 32.75
Target	regeste: Regeste a Schwere Bedrängnis, grosse seelische Belastung und heftige Gemütsbewegung; Art. 48 lit. a Ziff. 2 und Art. 48 lit. c StGB. Abgrenzung dieser Strafmilderungsgründe untereinander (E. 2.1-2.3). Regeste b Mord, heftige Gemütsbewegung und grosse seelische Belastung; Art. 48 lit. c und Art. 112 StGB. Die für den Mord typische Skrupellosigkeit schliesst die Bejahung der Strafmilderungsgründe der heftigen Gemütsbewegung und der grossen seelischen Belastung aus (E. 2.5).
Predicted	regeste: Regeste Art. 112 und 49 Abs. 2 StGB ; tiefe Gewalt, Schwere Gewalt, Mord. Die Beeinträchtigung oder der schwere Gewalt ist einen psychischen Klumpen im Sinne des Art. 48 StGB, wenn beide Personen einen Beziehungskonflikt bilden und sich ein bestimmtes Verhältnis zwischen einseitigen Umständen des Tatvergehens und der Begründung der Tatbegehren haben (E. 2 und 3).
Scores	BERT: 69.3, BLEU: 12.14, METEOR: 14.72, ROUGE-1/2/L: 30.43 / 10.26 / 30.43