Magglinger
Rechtsinformatikseminar

Forschungsprojekt «Open Justice vs. Privacy» und die Weiterentwicklung von Legal NLP

Daniel Kettiger, Kompetenzzentrum für Public Management der Universität Bern

Prof. Dr. Matthias Stürmer, Institut Public Sector Transformation der Berner Fachhochschule

eJustice.ch · Postfach · 3001 Bern · +41 58 462 47 23 · info@ejustice.ch

Agenda

1. Ergebnisse des Forschungsprojekts aus juristischer und verwaltungswissenschaftlicher Sicht
2. Ergebnisse des Forschungsprojekts aus technischer Sicht
3. Zukunft von Legal NLP (Natural Language Processing)

Forschungsprojekt «Open Justice vs. Privacy»

Forschungsprojekt im SNF NFP 77 «Digitale Transformation»

77 NFP Digitale Transformation
Nationales Forschungsprogramm

- Interdisziplinär: Recht, Computer Science, Verwaltungswissenschaften
- Transdisziplinär: Einbezug von Akteuren rund um die Gerichtsbarkeit

Universität Bern

- Kompetenzzentrum für Public Management (Prof. Dr. Andreas Lienhard)
- Institut für Wirtschaftsinformatik (Prof. Dr. Thomas Myrach)
- Institut für Informatik (Prof. Dr. Matthias Stürmer)
  + Berner Fachhochschule, Institut Public Sector Transformation

Forschungsgegenstand

- Anonymisierung von Gerichtsurteilen

https://www.nfp77.ch/de/c1Qw3hiAvSYfVs3z/projekt/gerichtsurteile-im-spannungsfeld-zwischen-transparenz-und-privatsphaere

Magglinger
Rechtsinformatikseminar

1. Ergebnisse des Forschungsprojekts aus juristischer und verwaltungswissenschaftlicher Sicht

77 NRP

Digital Transformation
National Research Programme

eJustice.ch · Postfach · 3001 Bern · +41 58 462 47 23 · info@ejustice.ch

Magglinger Rechtsinformatikseminar

Arbeiten im Rahmen des Forschungsprojekts

Juristische Forschung

- Dissertation «Publikation von Gerichtsurteilen im Spannungsfeld zwischen Transparenz und Geheimhaltungsinteressen», inkl. Befragung Gerichte CH und Rechtsvergleich Europa
- Masterarbeit: «Staatshaftung für mangelhafte Anonymisierung von publizierten Gerichtsurteilen » (Justice - Justiz - Giustizia 2022/1)
- Masterarbeit: «Datenschutzrechtliche Aspekte bei der Veröffentlichung von Gerichtsurteilen in sekundären Entscheid-Datenbanken (Justice - Justiz - Giustizia 2024/2)

Verwaltungswissenschaftliche Forschung

- Expertenbefragung zu Aspekten der Anonymisierung von Gerichtsurteilen
- Expertenworkshop

Fazit des NFP77-Projekts

Einige **Schlussfolgerungen und Empfehlungen** aus unserem Forschungsprojekt:

77 NFP **Digitale Transformation**
Nationales Forschungsprogramm

1. Der **Begriff der Anonymisierung** im Gerichts-organisations- und Verfahrensrecht entspricht nicht demjenigen im Datenschutzrecht.
2. Der «interne» Grund für die Anonymisierung von Gerichtsurteilen ist die **Vulnerabilität (Verletzlichkeit) der betroffenen Personen**. Besonders schutzbedürftige Personen (z. B. Kinder, Behinderte) benötigen bei der Veröffentlichung von Urteilen einen besonders sorgfältigen Persönlichkeitsschutz.
3. Das Urteil muss in jedem Fall **nachvollziehbar** bleiben, auch wenn nicht ausgeschlossen ist, dass eine mit den Einzelheiten des Falles bereits vertraute Person den Namen einer Partei erkennen könnte.

Magglinger
Rechtsinformatikseminar

2. Ergebnisse des Forschungsprojekts aus technischer Sicht

77 NRP

Digital Transformation
National Research Programme

eJustice.ch · Postfach · 3001 Bern · +41 58 462 47 23 · info@ejustice.ch

Grundkonzepte moderner KI-Modelle

Pre-Training

Generierung von
«Foundation Models»

→ Viele Daten (grosse Bild- und Textmengen)
Viel Rechenkapazität (GPUs)
Viel Energie (MWh)

Fine-Tuning

Spezialisierung für bestimmte Aufgaben, «Alignment»
→ kleine Datenmengen in hoher Qualität

Inference

Produktive Anwendung im regulären Betrieb
→ Prompting

Magglinger Rechtsinformatikseminar

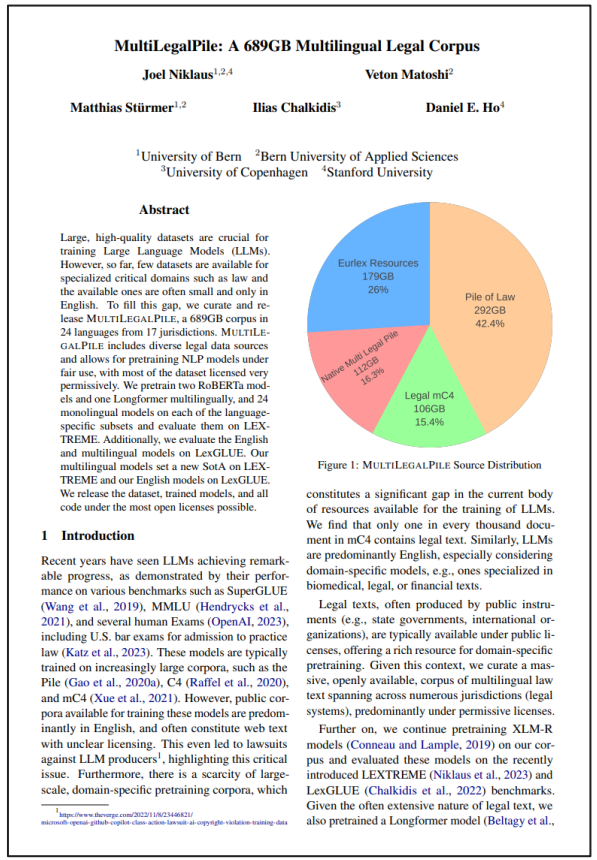
Mehrsprachige Rechtssprachen-Modell fürs Bundesgericht

Neues, juristisches Sprachmodell, das 2022 in einem Rechenzentrum von Google (TPU) auf Rechtstexten trainiert wurde **(Pre-Training)**

Auftrag des **Schweizerischen Bundesgerichts** zur besseren Anonymisierung von Gerichtsentscheiden

Dazu notwendig: Erstellung der **weltweit grössten mehrsprachigen Sammlung juristischer Texte:**

- Native Multi Legal Pile (112 GB)
- Eurlex-Ressourcen (179 GB)
- Legal MC4 (106 GB)
- Pile of Law (292 GB)



MultiLegalPile: A 689GB Multilingual Legal Corpus

Joel Niklaus[1,2,4] **Veton Matoshi**[2]
Matthias Stürmer[1,2] **Ilias Chalkidis**[3] **Daniel E. Ho**[4]
[1]University of Bern [2]Bern University of Applied Sciences
[3]University of Copenhagen [4]Stanford University

Abstract

Large, high-quality datasets are crucial for training Large Language Models (LLMs). However, so far, few datasets are available for specialized critical domains such as law and the available ones are often small and only in English. To fill this gap, we curate and release MULTILEGALPILE, a 689GB corpus in 24 languages from 17 jurisdictions. MULTILEGALPILE includes diverse legal data sources and allows for pretraining NLP models under fair use, with most of the dataset licensed very permissively. We pretrain two RoBERTa models and one Longformer multilingually, and 24 monolingual models on each of the language-specific subsets and evaluate them on LEXTREME. Additionally, we evaluate the English and multilingual models on LexGLUE. Our multilingual models set a new SotA on LEXTREME and our English models on LexGLUE. We release the dataset, trained models, and all code under the most open licenses possible.

Figure 1: MULTILEGALPILE Source Distribution

1 Introduction

Recent years have seen LLMs achieving remarkable progress, as demonstrated by their performance on various benchmarks such as SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), and several human Exams (OpenAI, 2023), including U.S. bar exams for admission to practice law (Katz et al., 2023). These models are typically trained on increasingly large corpora, such as the Pile (Gao et al., 2020a), C4 (Raffel et al., 2020), and mC4 (Xue et al., 2021). However, public corpora available for training these models are predominantly in English, and often constitute web text with unclear licensing. This even led to lawsuits against LLM producers[1], highlighting this critical issue. Furthermore, there is a scarcity of large-scale, domain-specific pretraining corpora, which constitutes a significant gap in the current body of resources available for the training of LLMs. We find that only one in every thousand document in mC4 contains legal text. Similarly, LLMs are predominantly English, especially considering domain-specific models, e.g., ones specialized in biomedical, legal, or financial texts.

Legal texts, often produced by public instruments (e.g., state governments, international organizations), are typically available under public licenses, offering a rich resource for domain-specific pretraining. Given this context, we curate a massive, openly available, corpus of multilingual law text spanning across numerous jurisdictions (legal systems), predominantly under permissive licenses.

Further on, we continue pretraining XLM-R models (Conneau and Lample, 2019) on our corpus and evaluated these models on the recently introduced LEXTREME (Niklaus et al., 2023) and LexGLUE (Chalkidis et al., 2022) benchmarks. Given the often extensive nature of legal text, we also pretrained a Longformer model (Beltagy et al.,

[1] https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

https://www.bfh.ch/en/research/research-projects/2021-432-033-021/

https://arxiv.org/abs/2306.02069

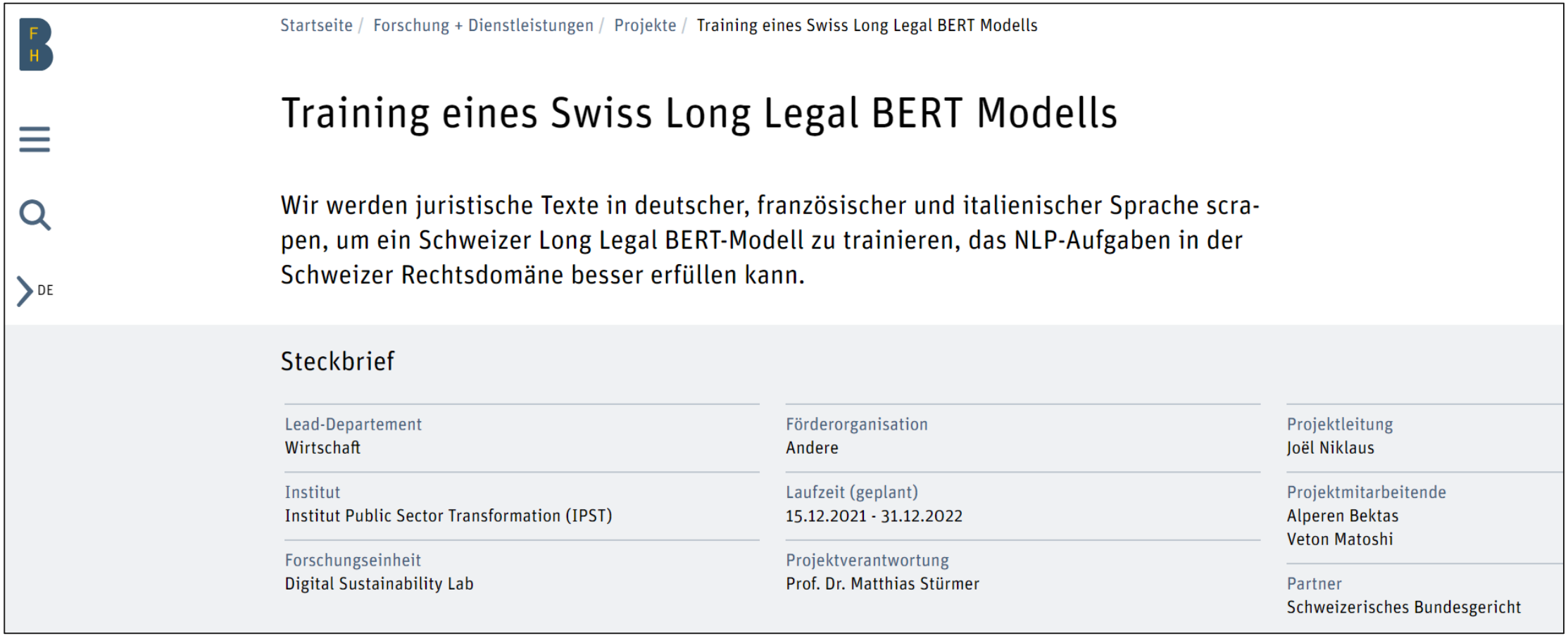
Weltweit grösste, mehrsprachige Rechtstextesammlung

Figure 2. MULTILEGALPILE Text Type Distribution

Language	Text Type	Words	Documents	Words per Document	Jurisdiction	Source	License/Copyright
Native Multi Legal Pile							
bg	legislation	309M	262k	1178	Bulgaria	MARCELL	CC0-1.0
cs	caselaw	571M	342k	1667	Czechia Czechia Czechia	CzCDC Constitutional Court CzCDC Supreme Administrative Court CzCDC Supreme Court	CC BY-NC 4.0 CC BY-NC 4.0 CC BY-NC 4.0
da	caselaw	211M	92k	2275	Denmark	DDSC	CC BY 4.0 and other, depending on the dataset
da	legislation	653M	296k	2201	Denmark	DDSC	CC BY 4.0 and other, depending on the dataset
de	caselaw	1786M	614k	2905	Germany Switzerland	openlegaldata entscheidsuche	ODbL-1.0 similar to CC BY
de	legislation	513M	302k	1698	Germany Switzerland	openlegaldata lexfind	ODbL-1.0 not protected by copyright law
en	legislation	2539M	713k	3557	Switzerland UK	lexfind uk-lex	not protected by copyright law CC BY 4.0
fr	caselaw	1172M	495k	2363	Belgium France Luxembourg Switzerland	jurportal CASS judoc entscheidsuche	not protected by copyright law Open Licence 2.0 not protected by copyright law similar to CC BY
fr	legislation	600M	253k	2365	Switzerland Belgium	lexfind ejustice	not protected by copyright law not protected by copyright law
hu	legislation	265M	259k	1019	Hungary	MARCELL	CC0-1.0
it	caselaw	407M	159k	2554	Switzerland	entscheidsuche	similar to CC BY
it	legislation	543M	238k	2278	Switzerland	lexfind	not protected by copyright law
nl	legislation	551M	243k	2263	Belgium	ejustice	not protected by copyright law
pl	legislation	299M	260k	1148	Poland	MARCELL	CC0-1.0
pt	caselaw	12613M	17M	728	Brazil Brazil Brazil	RulingBR CRETA CJPG	not protected by copyright law CC BY-NC-SA 4.0 not protected by copyright law
ro	legislation	559M	396k	1410	Romania	MARCELL	CC0-1.0
sk	legislation	280M	246k	1137	Slovakia	MARCELL	CC0-1.0
sl	legislation	366M	257k	1418	Slovenia	MARCELL	CC-BY-4.0
total		**24236M**	**23M**	**1065**		**Native Multi Legal Pile**	
Overall statistics for the remaining subsets							
total		**12107M**	**8M**	**1457**	EU	**Eurlex Resources**	**CC BY 4.0**
total		**43376M**	**18M**	**2454**	US (99%), Canada, and EU	**Pile of Law**	CC BY-NC-SA 4.0; **See Henderson* et al. (2022) for details**
total		**28599M**	**10M**	**2454**		**Legal mC4**	**ODC-BY**

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024 “MultiLegalPile: A 689GB Multilingual Legal Corpus” Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics https://aclanthology.org/2024.acl-long.805/

Neues Rechts-Sprachmodell für Bundesgericht trainiert



Startseite / Forschung + Dienstleistungen / Projekte / Training eines Swiss Long Legal BERT Modells

Training eines Swiss Long Legal BERT Modells

Wir werden juristische Texte in deutscher, französischer und italienischer Sprache scrapen, um ein Schweizer Long Legal BERT-Modell zu trainieren, das NLP-Aufgaben in der Schweizer Rechtsdomäne besser erfüllen kann.

DE

Steckbrief

Lead-Departement
Wirtschaft

Institut
Institut Public Sector Transformation (IPST)

Forschungseinheit
Digital Sustainability Lab

Förderorganisation
Andere

Laufzeit (geplant)
15.12.2021 - 31.12.2022

Projektverantwortung
Prof. Dr. Matthias Stürmer

Projektleitung
Joël Niklaus

Projektmitarbeitende
Alperen Bektas
Veton Matoshi

Partner
Schweizerisches Bundesgericht

https://www.bfh.ch/de/forschung/forschungsprojekte/2021-432-033-021/

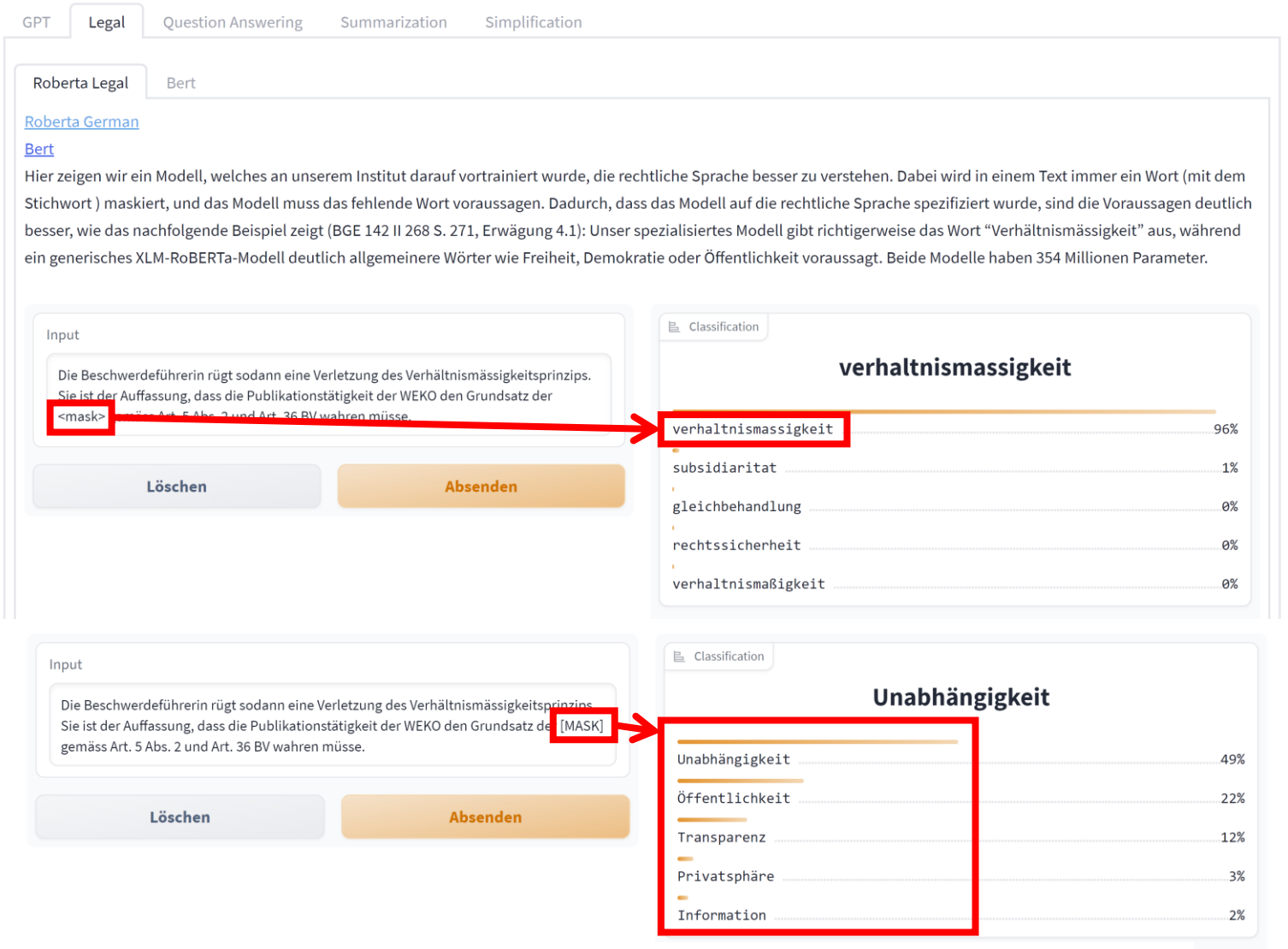
Vergleich

Domain-spezifisches Modell versteht Rechtssprache besser als generische Modell.

Beispiel: das Wort «**Verhältnismässigkeit**» ist maskiert («MASK»)

Juristisches Rechtssprachen-Modell findet Wort sofort (96% Wahrscheinlichkeit)

Generisches Sprachmodell vermutet «Unabhängigkeit» (49%), «Öffentlichkeit» (22%) oder «Transparenz» (12%)



https://www.bfh.ch/wirtschaft/de/forschung/forschungsbereiche/nlp-demo/

Magglinger Rechtsinformatikseminar

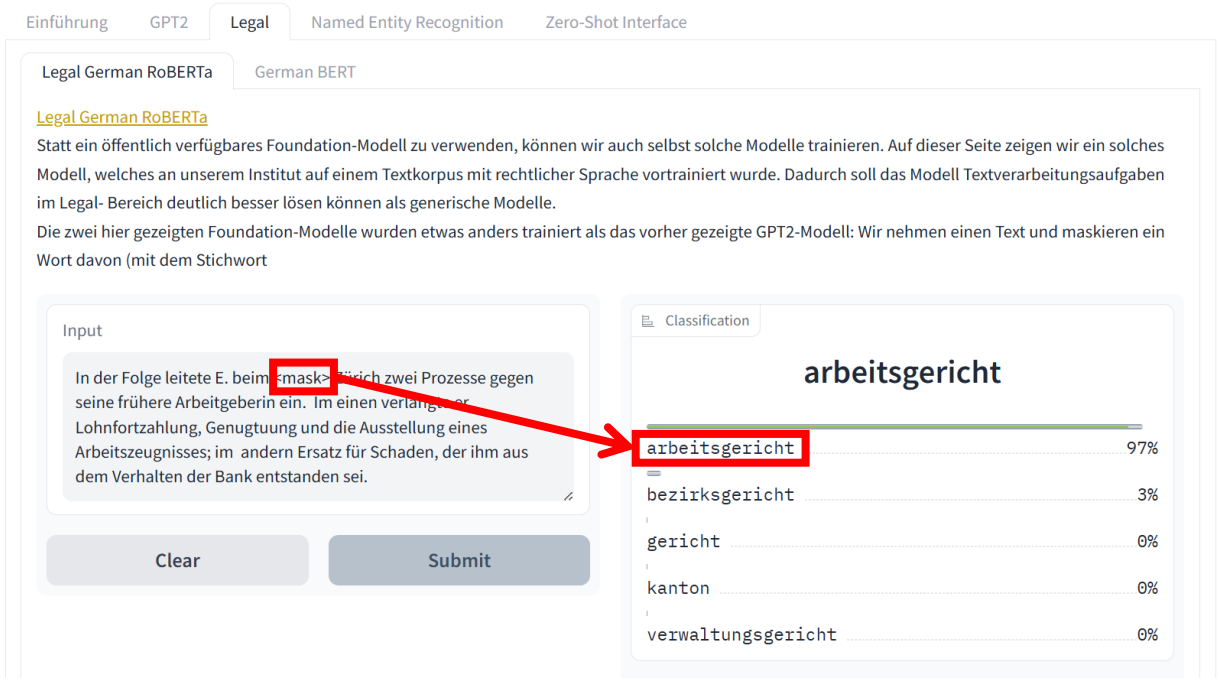
Vergleich mit ChatGPT

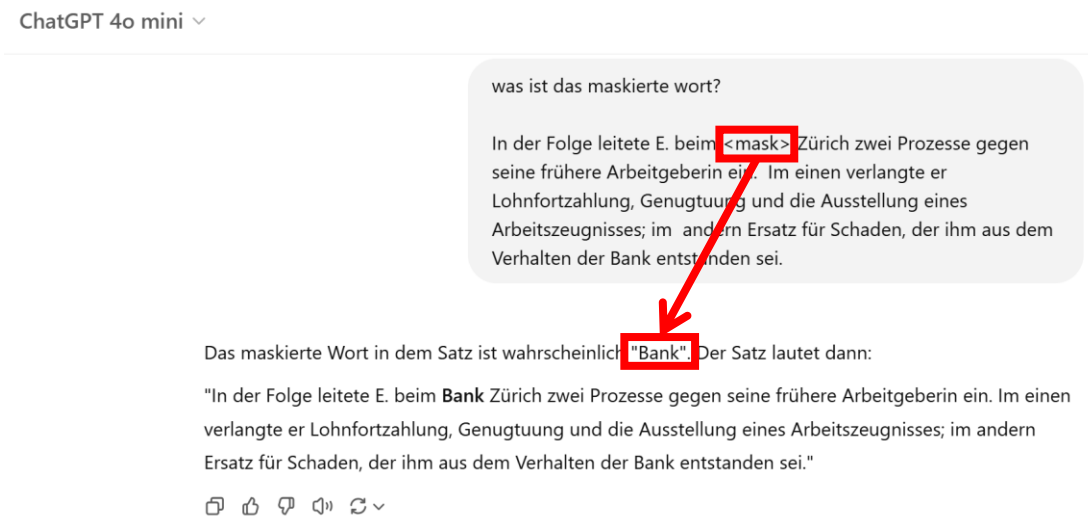
Kleine, Domain-spezifische Modelle können sogar besser als ChatGPT sein!

Beispiel: das Wort **«Arbeitsgericht»** ist maskiert («MASK»)

Juristisches Rechtssprachen-Modell findet Wort sofort (97% Wahrscheinlichkeit)

ChatGPT vermutet «Bank»…



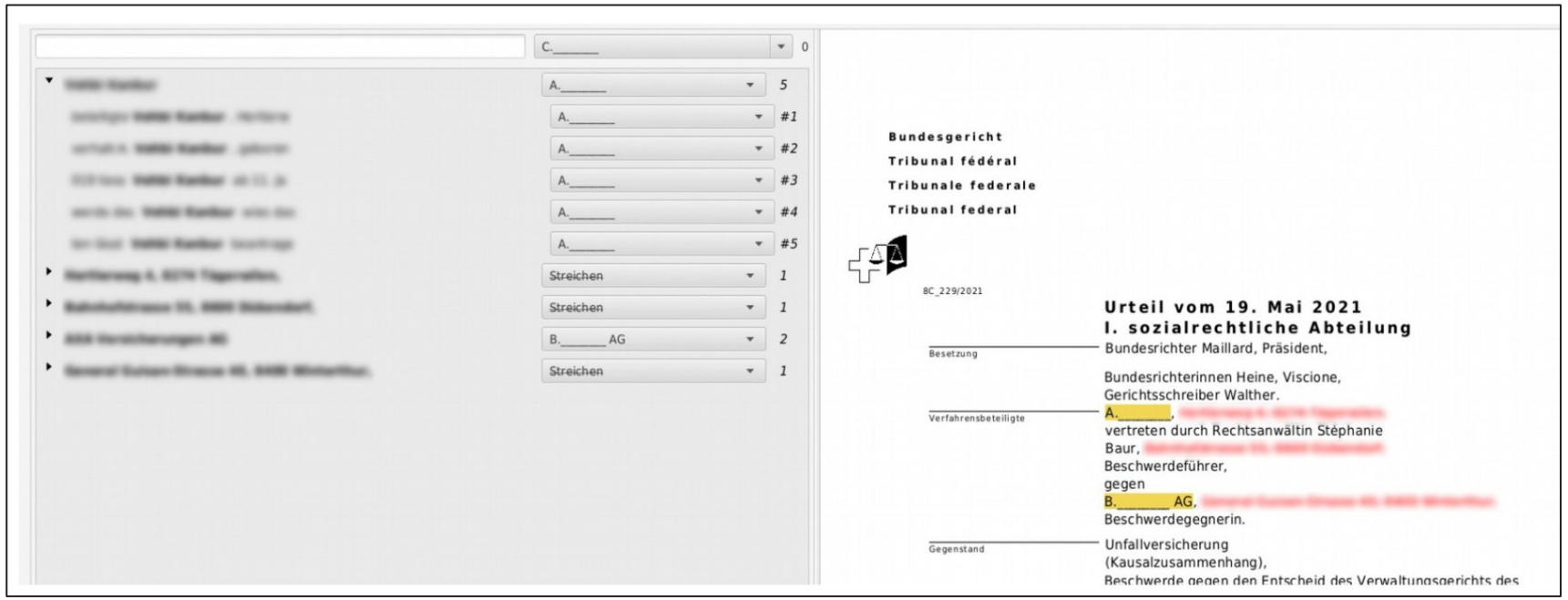


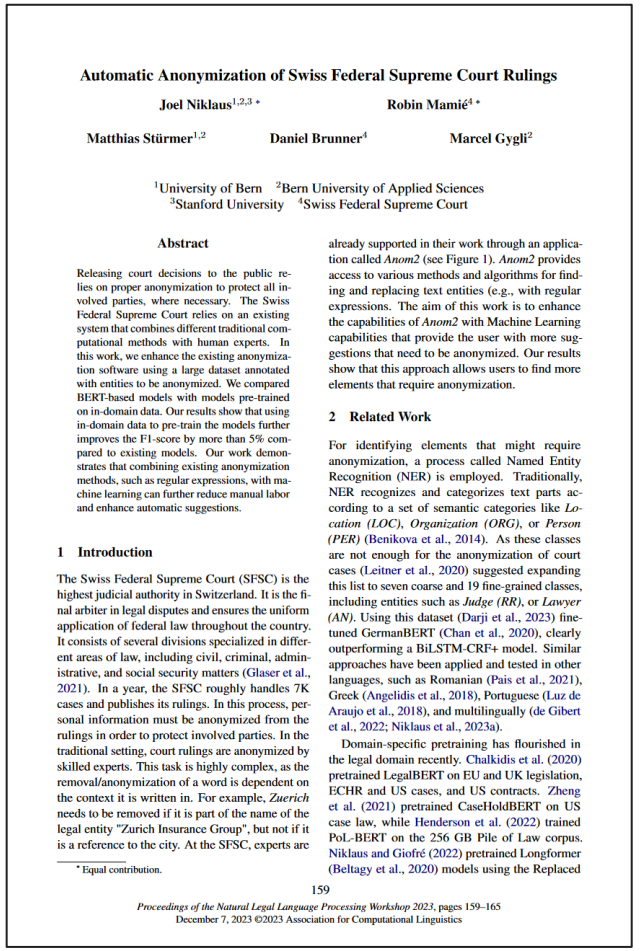
https://www.bfh.ch/wirtschaft/de/forschung/forschungsbereiche/nlp-demo/

Magglinger
Rechtsinformatikseminar

Verbesserung der Anonymisierung mittels NLP

Betrieb von KI-Modell durch **interne Informatikabteilung des Schweizer Bundesgerichts** für Anonymisierungs-Vorschläge





Automatic Anonymization of Swiss Federal Supreme Court Rulings

Joel Niklaus[1,2,3 *] Robin Mamié[4 *]
Matthias Stürmer[1,2] Daniel Brunner[4] Marcel Gygli[2]
[1]University of Bern [2]Bern University of Applied Sciences
[3]Stanford University [4]Swiss Federal Supreme Court

Abstract

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

1 Introduction

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

2 Related Work

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

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

* Equal contribution.

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

Joel Niklaus, Robin Mamié, Matthias Stürmer, Daniel Brunner, und Marcel Gygli 2023 „Automatic Anonymization of Swiss Federal Supreme Court Rulings“.
In Proceedings of the Natural Legal Language Processing Workshop 2023, 159–65. Singapore: Association for Computational Linguistics, 2023.
https://doi.org/10.18653/v1/2023.nllp-1.16

Prüfen der De-Anonymisierungs-Risiken mit LLMs

Experimente mit der Re-Identifizierung von Gerichtsurteilen und Wikipedia-Seiten

Testen verschiedener De-Anonymisierungs-Tasks:

- „Maske ausfüllen“
- „Fragenbeantwortung“
- „Text-Erzeugung“

Erstellung von Metriken zur Messung der De-Anonymisierung:

- Partial Name Match Score (PNMS)
- Normalisierte Levenshtein-Distanz (NLD)
- Ergebnis der Nachnamenübereinstimmung (LNMS)
- Gewichtete Teilnamensübereinstimmung (W-PNMS)

Anonymity at Risk?
Assessing Re-Identification Capabilities of Large Language Models

Alex Nyffenegger[1,2*] Matthias Stürmer [2,3] Joel Niklaus [2,3,4*]
[1]University of Fribourg
[2]University of Bern
[3]Bern University of Applied Sciences
[4]Stanford University

Abstract

Anonymity of both natural and legal persons in court rulings is a critical aspect of privacy protection in the European Union and Switzerland. With the advent of LLMs, concerns about large-scale re-identification of anonymized persons are growing. In accordance with the Federal Supreme Court of Switzerland, we explore the potential of LLMs to re-identify individuals in court rulings by constructing a proof-of-concept using actual legal data from the Swiss federal supreme court. Following the initial experiment, we constructed an anonymized Wikipedia dataset as a more rigorous testing ground to further investigate the findings. With the introduction and application of the new task of re-identifying people in texts, we also introduce new metrics to measure performance. We systematically analyze the factors that influence successful re-identifications, identifying model size, input length, and instruction tuning among the most critical determinants. Despite high re-identification rates on Wikipedia, even the best LLMs struggled with court decisions. The complexity is attributed to the lack of test datasets, the necessity for substantial training resources, and data sparsity in the information used for re-identification. In conclusion, this study demonstrates that re-identification using LLMs may not be feasible for now, but as the proof-of-concept on Wikipedia showed, it might become possible in the future. We hope that our system can help enhance the confidence in the security of anonymized decisions, thus leading to the courts being more confident to publish decisions.

Figure 1: Re-identification rate by parameter count

1 Introduction

The swift advancements in Natural Language Processing (NLP) (Vaswani et al. 2017; Brown et al. 2020; Ouyang et al. 2022; Khurana et al. 2023) have introduced new challenges to the security of traditional legal processes (Tsarapatsanis and Aletras 2021). As public access to data increases in tandem with digital advancements (Katz et al. 2023; EUGH 2018; Lorenz 2017), the potential risks associated with data disclosure have become increasingly significant. Increasingly larger and more capable Large Language Models (LLMs), more powerful vector stores and potent embeddings together have the capacity to extract unintended information from public data (Borgeaud et al. 2022; Carlini et al. 2021). This poses a security risk, as the identification of individuals involved in legal proceedings can lead to privacy breaches, providing undue advantage to certain legal actors, and risking public defamation. Over the past decade, at least 18 requests for name changes following the re-identification of convicts have been registered in Switzerland, indicating that this issue already exists due to imprudent media coverage (Stückelberger, Evin, and Damian 2021). The number of cases where the accused become victims of unlawful personal information disclosure is likely to rise as further re-identifications occur. The prevention of re-identification is critical not only for the protection of the accused, but also for the courts. Munz (2022) even suggests that the state could be held accountable for non-monetary damages to judged persons, underscoring the urgent need for courts to address the re-identification issue proactively. Vokinger and Mühlematter (2019) have shown that some re-identifications are possible by applying regular expressions.

We use state-of-the-art transformer models (Vaswani et al. 2017) like LLaMA-2, GPT-4 or BLOOM (Touvron et al. 2023; OpenAI 2023; Scao et al. 2023) to re-identify individuals in publicly released Swiss court decisions. Such models have the ability to store information within their parameters and extract this information when prompted (Roberts, Raffel, and Shazeer 2020; AlKhamissi et al. 2022; Ippolito et al. 2023; Carlini et al. 2023). We find that while the best models are capable of identifying persons from masked Wikipedia

*These authors contributed equally.

Nyffenegger et al. 2023 “Anonymity at Risk? Assessing Re-Identification Capabilities of LLMs” https://arxiv.org/abs/2308.11103

Fazit des NFP77-Projekts

77 NFP **Digitale Transformation**
Nationales Forschungsprogramm

Einige **Schlussfolgerungen und Empfehlungen** aus unserem Forschungsprojekt:

1. Zurzeit besteht **kein Risiko** der vollautomatisierten De-Anonymisierung von Gerichtsurteilen durch künstliche Intelligenz (KI)
2. Um künftige Re-Identifikation zu verhindern, wird der Einsatz von **KI-unterstützten Anonymisierungs-Tools** empfohlen
3. Das **technische Potenzial** und damit das Risiko zur Re-Identifizierung ist systematisch zu **überwachen** und periodisch zu evaluieren

Magglinger
Rechtsinformatikseminar

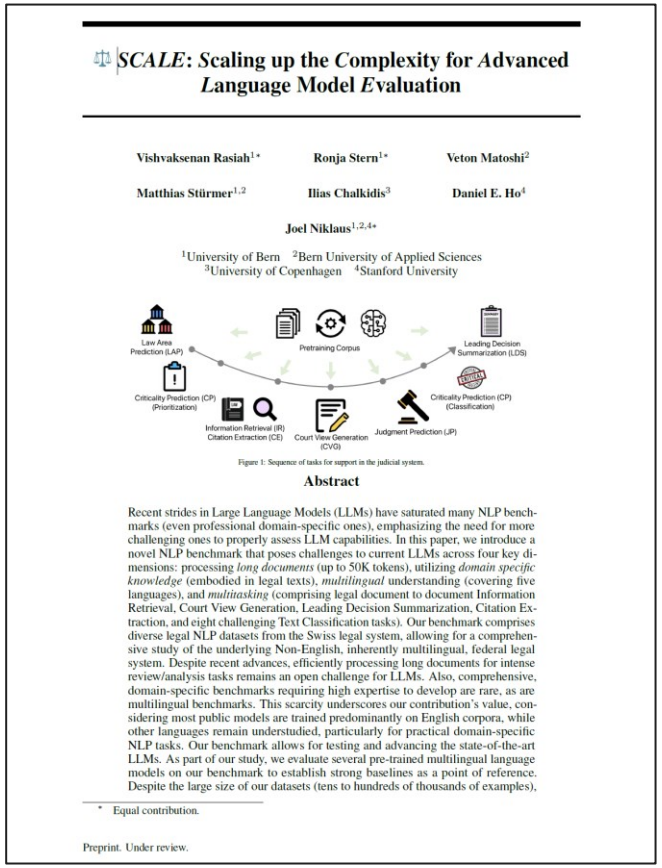
3. Zukunft von Legal NLP (Natural Language Processing)

eJustice.ch · Postfach · 3001 Bern · +41 58 462 47 23 · info@ejustice.ch

Benchmarks (Qualitätsmessung) typischer Gerichtsaufgaben

Neue Sammlung juristischer NLP-Benchmarks mit einzigartigen Merkmalen:

1. **Verarbeitung langer Dokumente**
   → bis zu 50’000 Tokens (Wortteile)
2. **Domänenspezifisches Wissen**
   → in juristischen Texten verankert
3. **Mehrsprachiges Textverständnis**
   → in fünf Sprachen (DE, FR, IT, RM, EN)
4. **Multitasking (unterschiedliche Aufgaben)**
   → Information Retrieval, Court View Generation, Leading Decision Summarization, Citation Extraction, Text Classification



SCALE: Scaling up the Complexity for Advanced Language Model Evaluation

Vishvaksenan Rasiah[1,*] Ronja Stern[1,*] Veton Matoshi[2]
Matthias Stürmer[1,2] Ilias Chalkidis[3] Daniel E. Ho[4]
Joel Niklaus[1,2,4,*]
[1]University of Bern [2]Bern University of Applied Sciences
[3]University of Copenhagen [4]Stanford University

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

Abstract

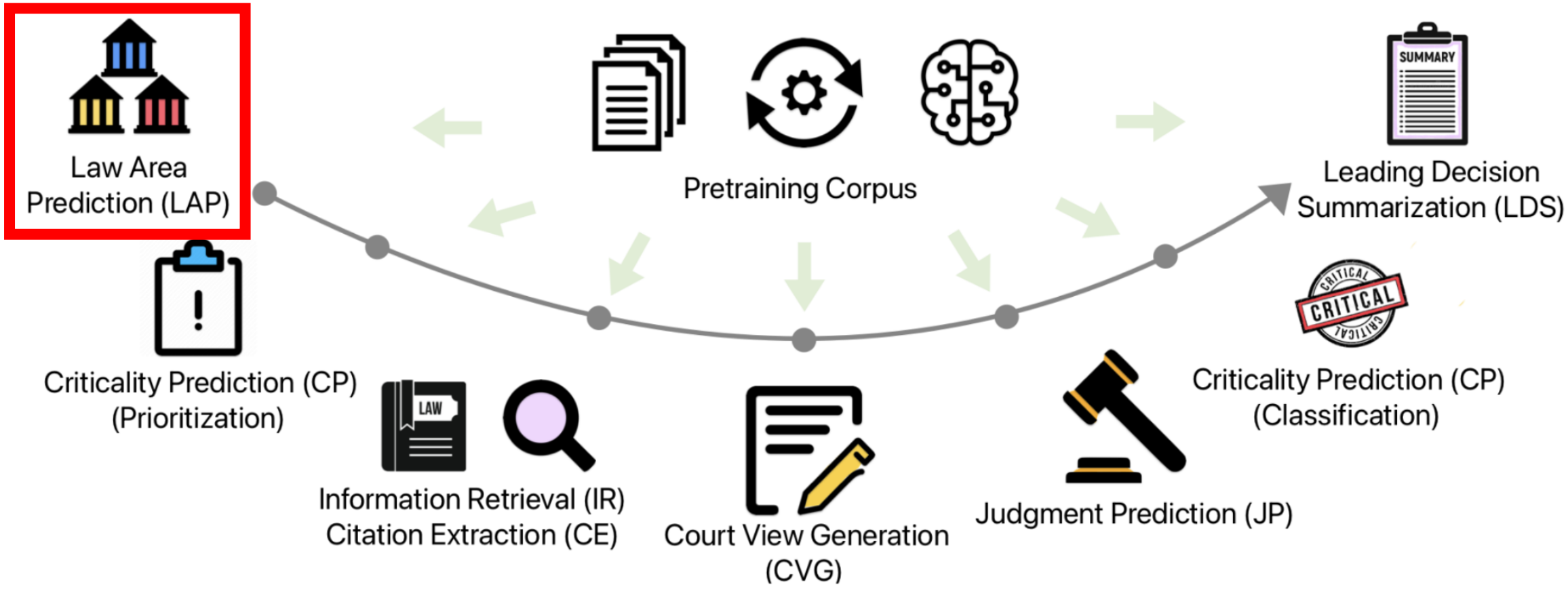
Recent strides in Large Language Models (LLMs) have saturated many NLP benchmarks (even professional domain-specific ones), emphasizing the need for more challenging ones to properly assess LLM capabilities. In this paper, we introduce a novel NLP benchmark that poses challenges to current LLMs across 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 LLMs. 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 (tens to hundreds of thousands of examples),

* Equal contribution.

Preprint. Under review.

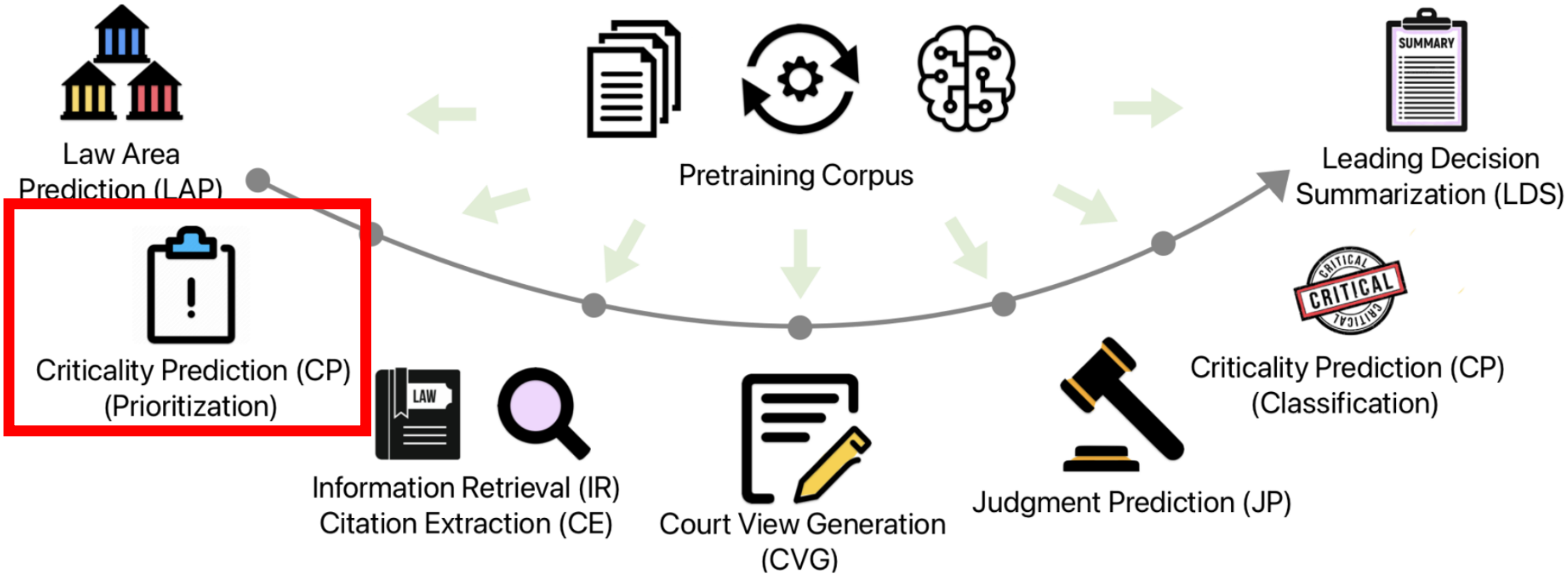
Rasiah et al. 2023 “SCALE: Scaling up the Complexity for Advanced Language Model Evaluation” http://arxiv.org/abs/2306.09237

Legal NLP für Gerichte: 1) Vorhersage des Rechtsgebiets



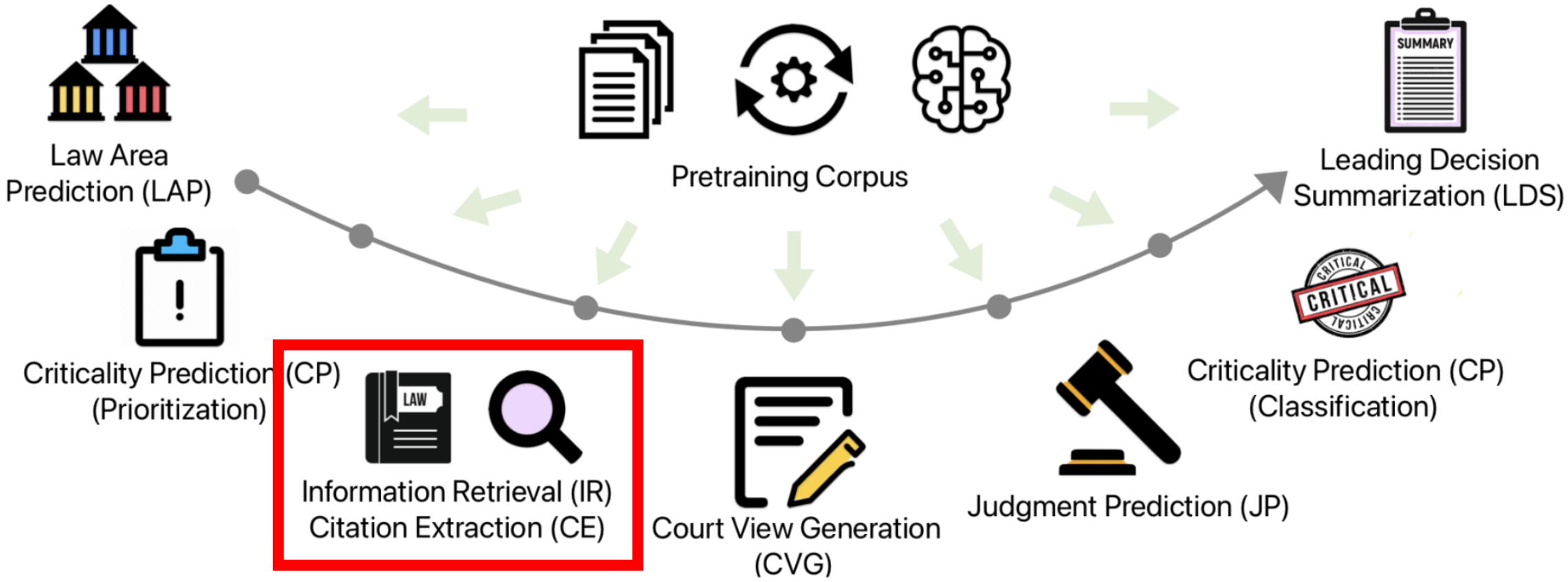
Stern et al. 2024 “One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support” http://arxiv.org/abs/2306.09237

Legal NLP für Gerichte: 2) Vorhersage der Relevanz (Priorität)



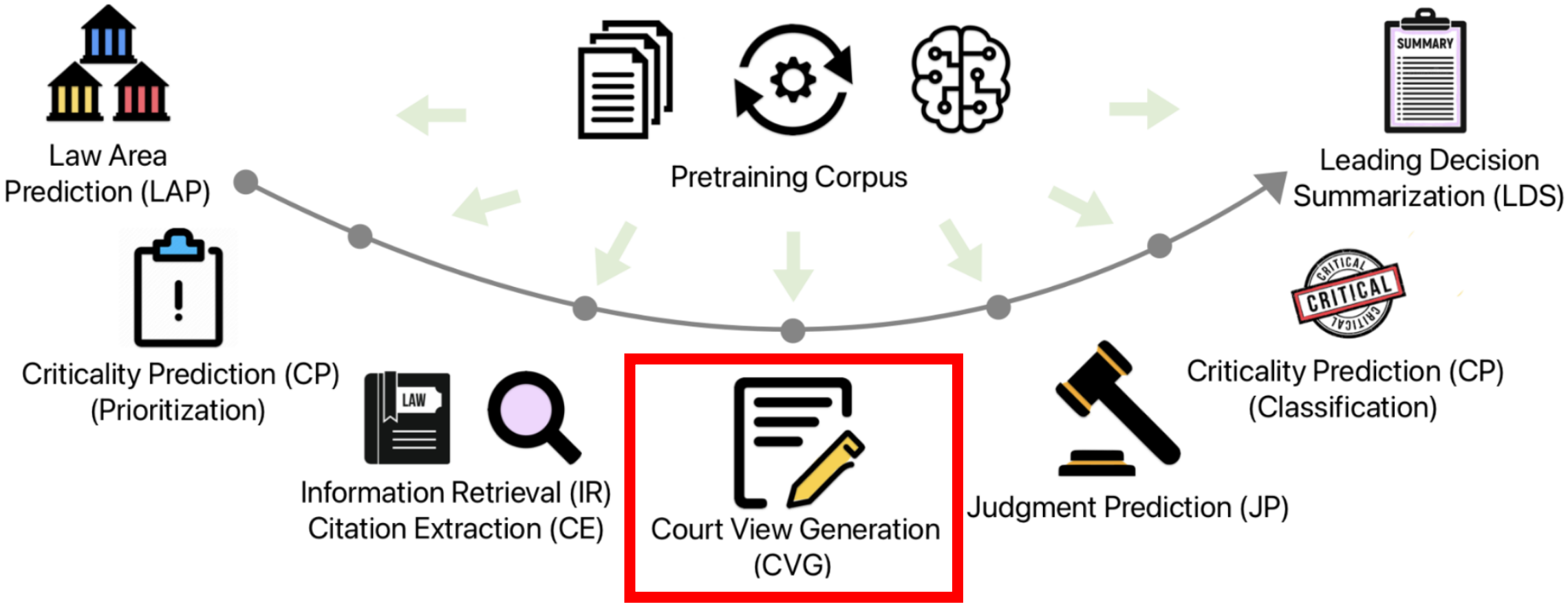
Stern et al. 2024 “One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support” http://arxiv.org/abs/2306.09237

Legal NLP für Gerichte: 3) Informations-Extraktion (Suche)



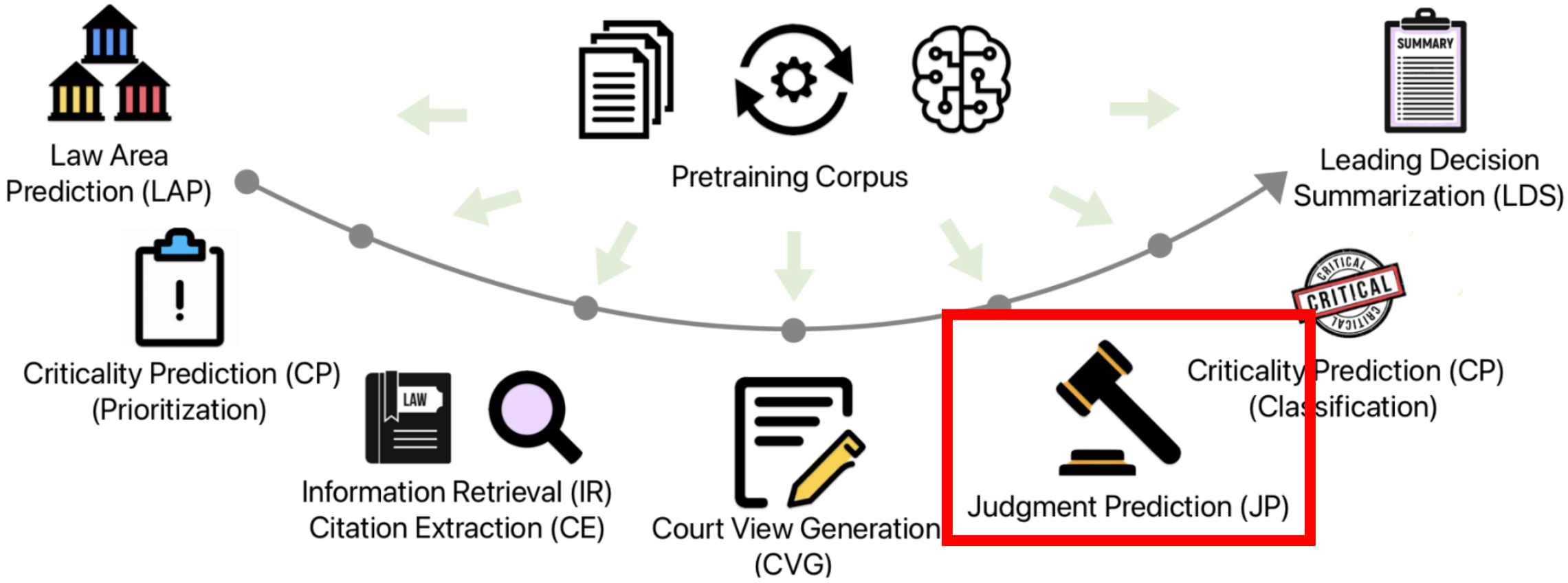
Stern et al. 2024 “One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support” http://arxiv.org/abs/2306.09237

Legal NLP für Gerichte: 4) Generierung der Gerichtsperspektive



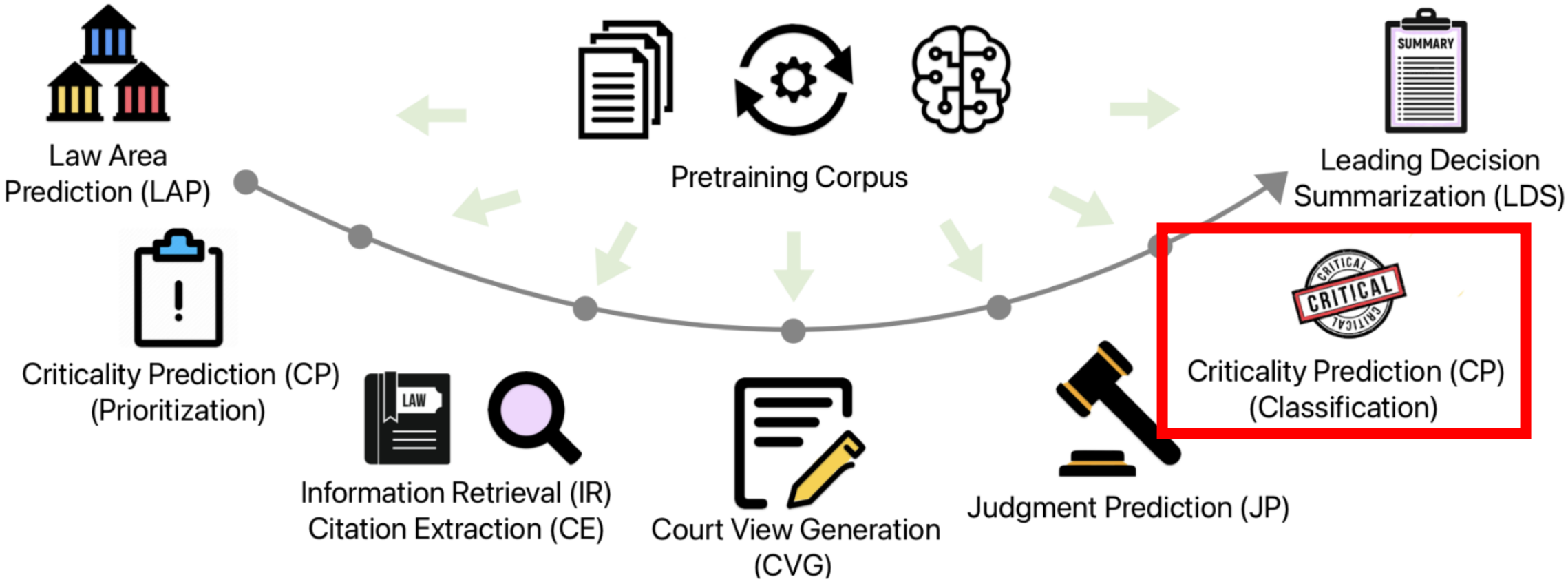
Stern et al. 2024 “One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support” http://arxiv.org/abs/2306.09237

Legal NLP für Gerichte: 5) Urteilsvorhersage



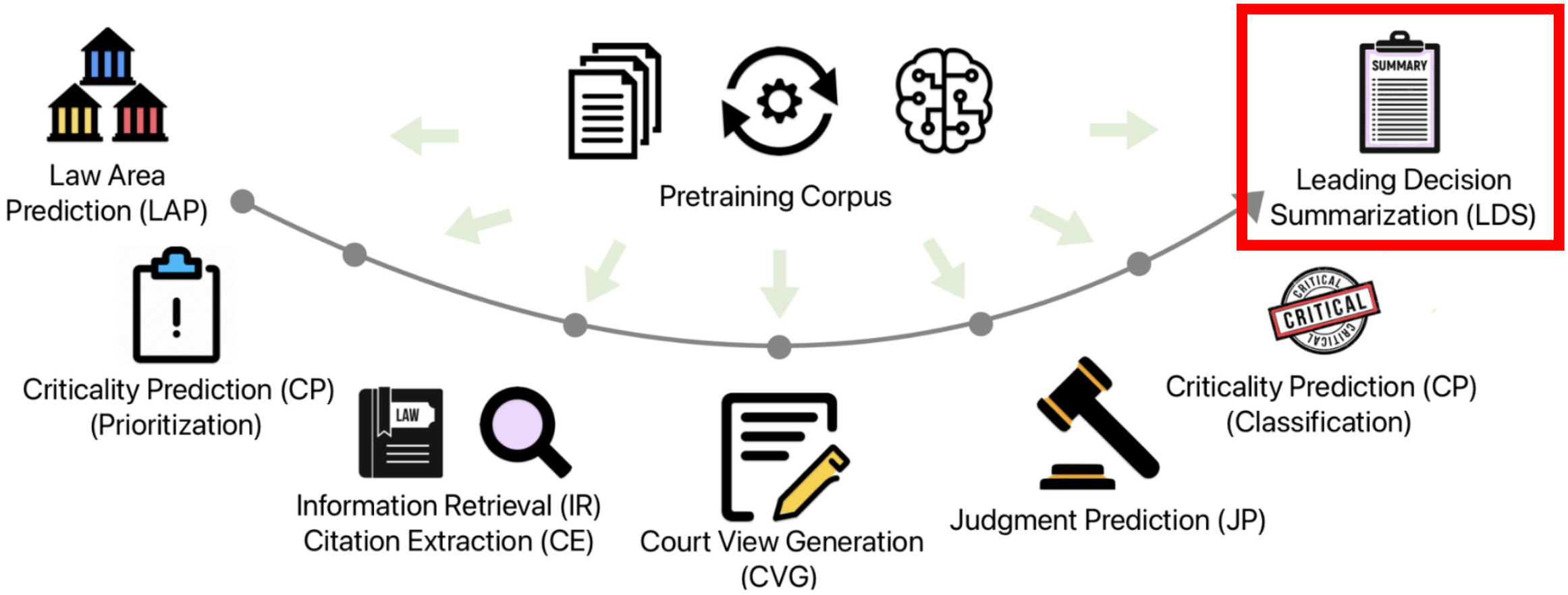
Stern et al. 2024 “One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support” http://arxiv.org/abs/2306.09237

Legal NLP für Gerichte: 6) Relevanz (Klassifizierung von Urteilen)



Stern et al. 2024 “One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support” http://arxiv.org/abs/2306.09237

Legal NLP für Gerichte: 7) Entscheidungszusammenfassung



Stern et al. 2024 “One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support” http://arxiv.org/abs/2306.09237

Legal Reasoning

- **Benchmark** (=Testing) von rechtlichen Schlussfolgerungen
- 162 Aufgaben in **6 Fachgebieten,** erstellt von Jurist:innen:
    1. Problemerkennung
    2. Regelaufruf
    3. Regelanwendung
    4. Regelschluss
    5. Interpretation
    6. rhetorisches Verständnis
- Prüfung von **20 Open Source und proprietären KI-Modellen**

LEGALBENCH: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models

Neel Guha[1], Julian Nyarko*[1], Daniel E. Ho*[1], Christopher Ré*[1], Adam Chilton[2], Aditya Narayana[3], Alex Chohlas-Wood[1], Austin Peters[1], Brandon Waldon[1], Daniel N. Rockmore[4], Diego Zambrano[1], Dmitry Talisman[3], Enam Hoque[5], Faiz Surani[1], Frank Fagan[6], Galit Sarfaty[7], Gregory M. Dickinson[8], Haggai Porat[9], Jason Hegland[1], Jessica Wu[1], Joe Nudell[1], Joel Niklaus[1], John Nay[10], Jonathan H. Choi[11], Kevin Tobia[12], Margaret Hagan[13], Megan Ma[10], Michael Livermore[14], Nikon Rasumov-Rahe[3], Nils Holzenberger[15], Noam Kolt[7], Peter Henderson[1], Sean Rehaag[16], Sharad Goel[17], Shang Gao[20], Spencer Williams[18], Sunny Gandhi[19], Tom Zur[9], Varun Iyer , and Zehua Li[1]

[1]Stanford University, [2]University of Chicago, [3]Maxime Tools, [4]Dartmouth College, [5]LawBeta, [6]South Texas College of Law Houston, [7]University of Toronto, [8]St. Thomas University Benjamin L. Crump College of Law, [9]Harvard Law School, [10]Stanford Center for Legal Informatics - CodeX, [11]University of Southern California, [12]Georgetown University Law Center, [13]Stanford Law School, [14]University of Virginia, [15]Télécom Paris, Institut Polytechnique de Paris, [16]Osgoode Hall Law School, York University, [17]Harvard Kennedy School, [18]Golden Gate University School of Law, [19]Luddy School of Informatics - Indiana University Bloomington, [20]Casetext

Abstract

The advent of large language models (LLMs) and their adoption by the legal community has given rise to the question: what types of legal reasoning can LLMs perform? To enable greater study of this question, we present LEGALBENCH: a collaboratively constructed legal reasoning benchmark consisting of 162 tasks covering six different types of legal reasoning. LEGALBENCH was built through an interdisciplinary process, in which we collected tasks designed and hand-crafted by legal professionals. Because these subject matter experts took a leading role in construction, tasks either measure legal reasoning capabilities that are practically useful, or measure reasoning skills that lawyers find interesting. To enable cross-disciplinary conversations about LLMs in the law, we additionally show how popular legal frameworks for describing legal reasoning—which distinguish between its many forms—correspond to LEGALBENCH tasks, thus giving lawyers and LLM developers a common vocabulary. This paper describes LEGALBENCH, presents an empirical evaluation of 20 open-source and commercial LLMs, and illustrates the types of research explorations LEGALBENCH enables.

1 Introduction

Advances in large language models (LLMs) are leading American legal professionals to reexamine the practice of law [52, 61, 158, 55, 13].[1] Proponents have argued that LLMs could alter how lawyers approach tasks ranging from brief writing to corporate compliance [158]. By making legal services more accessible, they could eventually help alleviate the United States' long standing access-to-justice crisis [35, 132]. This perspective is informed by the observation that LLMs possess special properties

[1]In using "LLMs", we are referring to language models which evince in-context learning capabilities (also referred to as "foundation models" [13]). This behavior has traditionally been observed in models with at least a billion parameters.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

Guha et al. 2024 "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models" 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks
https://proceedings.neurips.cc/paper_files/paper/2023/file/89e44582fd28ddfea1ea4dcb0ebbf4b0-Paper-Datasets_and_Benchmarks.pdf

eJustice.ch Verein Association Associazione

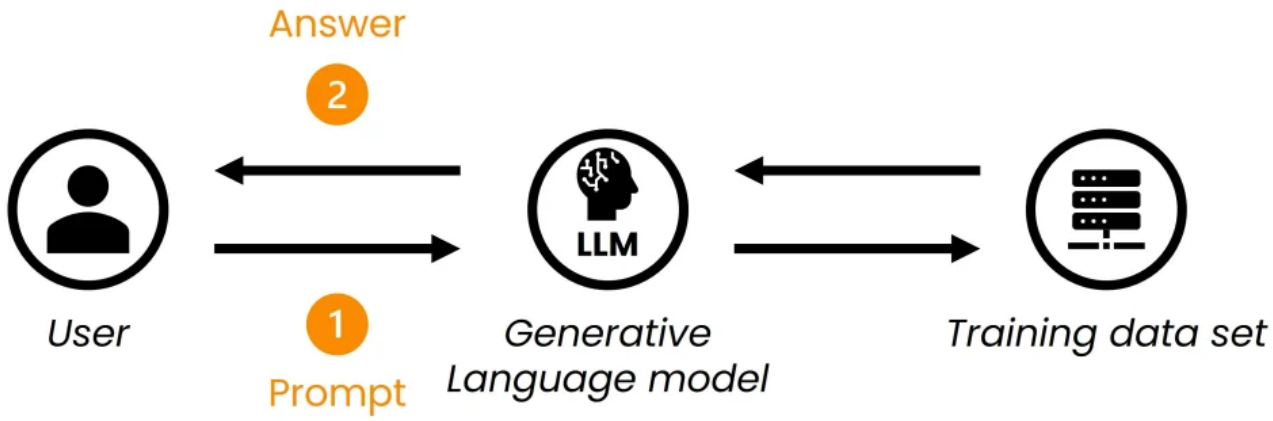
Vielen Dank für Ihre Aufmerksamkeit

eJustice.ch · Postfach · 3001 Bern · +41 58 462 47 23 · info@ejustice.ch

Magglinger Rechtsinformatikseminar

Backup Folien

Reguläres Prompting von einem LLM (bspw. ChatGPT)



Probleme:

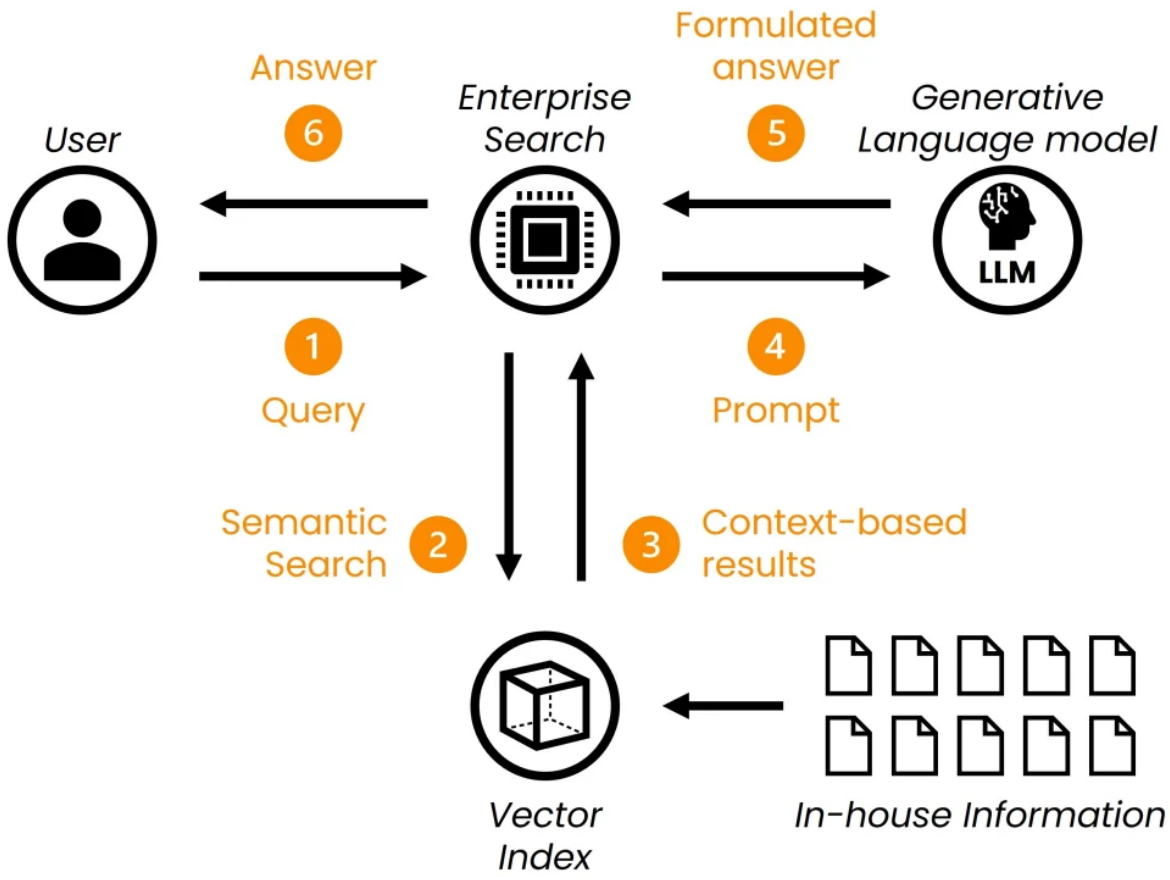
Kombiniert nach Wahrscheinlichkeit → halluziniert, nicht zuverlässig

Kann keine Fragen mit Wissen ausserhalb der Trainingsdaten beantworten, nur Fragen mit Informationen, die im Trainingsdatensatz enthalten sind

Kann keine Quellen zu den Antworten hinzufügen → unklar woher Informationen

Quelle: https://ambersearch.de/what-is-retrieval-augmented-generation

Retrieval Augmented Generation (RAG) Verfahren hat Potenzial



Vorteile von RAG:

- **Wenig Halluzinationen**
  → LLM dient nur zur Formulierung des Textes und erstellt keine neuen Inhalte
- **Quellenangaben** möglich
  → Herkunft der Informationen ist verfügbar
- **Einfache Aktualisierung** der Informationen → aktuelle Daten schnell verfügbar
- **Zugriffsrechte und Relevanz** → nur erlaubte Informationen zugänglich

Quelle: https://ambersearch.de/what-is-retrieval-augmented-generation