ONE LAW, MANY LANGUAGES: BENCHMARKING MULTILINGUAL LEGAL REASONING FOR JUDICIAL SUPPORT

Vishvaksenan Rasiah[1,*], Ronja Stern[1,*], Veton Matoshi[2], Matthias Stürmer[1,2], Ilias Chalkidis[3], Daniel E.Ho[4], Joël Niklaus[1,2,*]

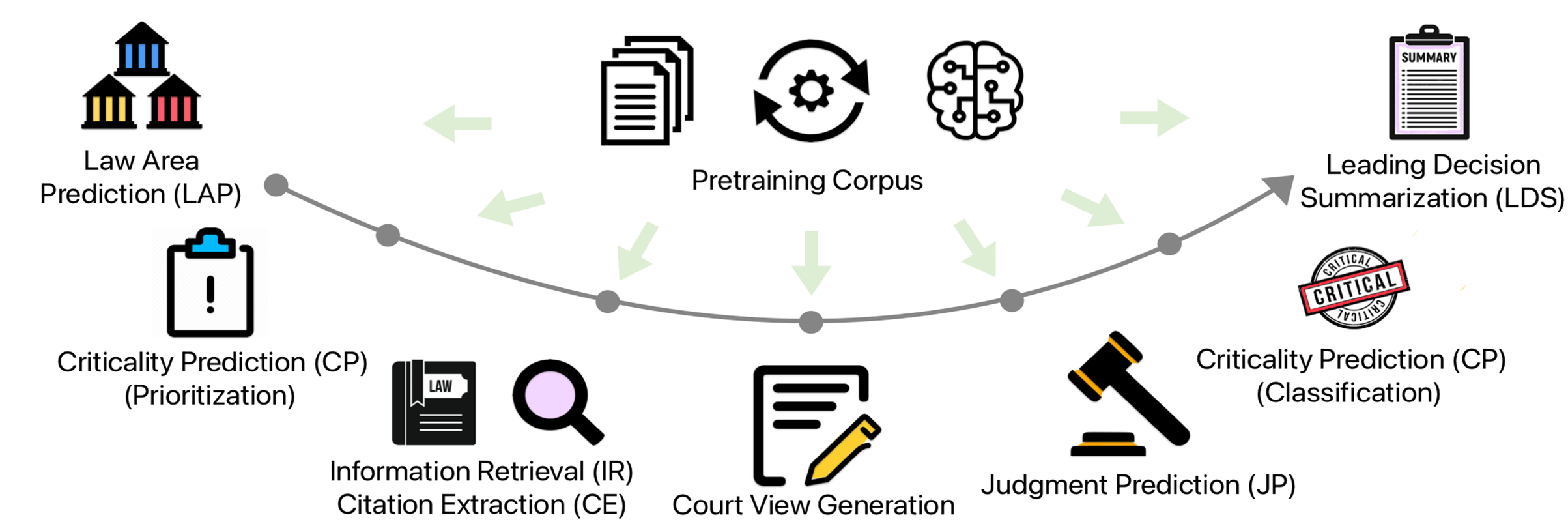
[1]University of Bern, [2]Bern University of Applied Sciences, [3]University of Copenhagen, [4]Stanford University

Main Contributions:

- Seven public multilingual datasets containing Swiss legal documents
- Release of two large in-domain pre-training datasets and three pre-trained models
- Evaluate multilingual language models on our benchmark and compare them to our models

Challenges:

- Processing long documents
- Using domain-specific knowledge
- Multilingualism
- Multitasking (Information Retrival, Text Generation and Classification)
- Legal reasoning



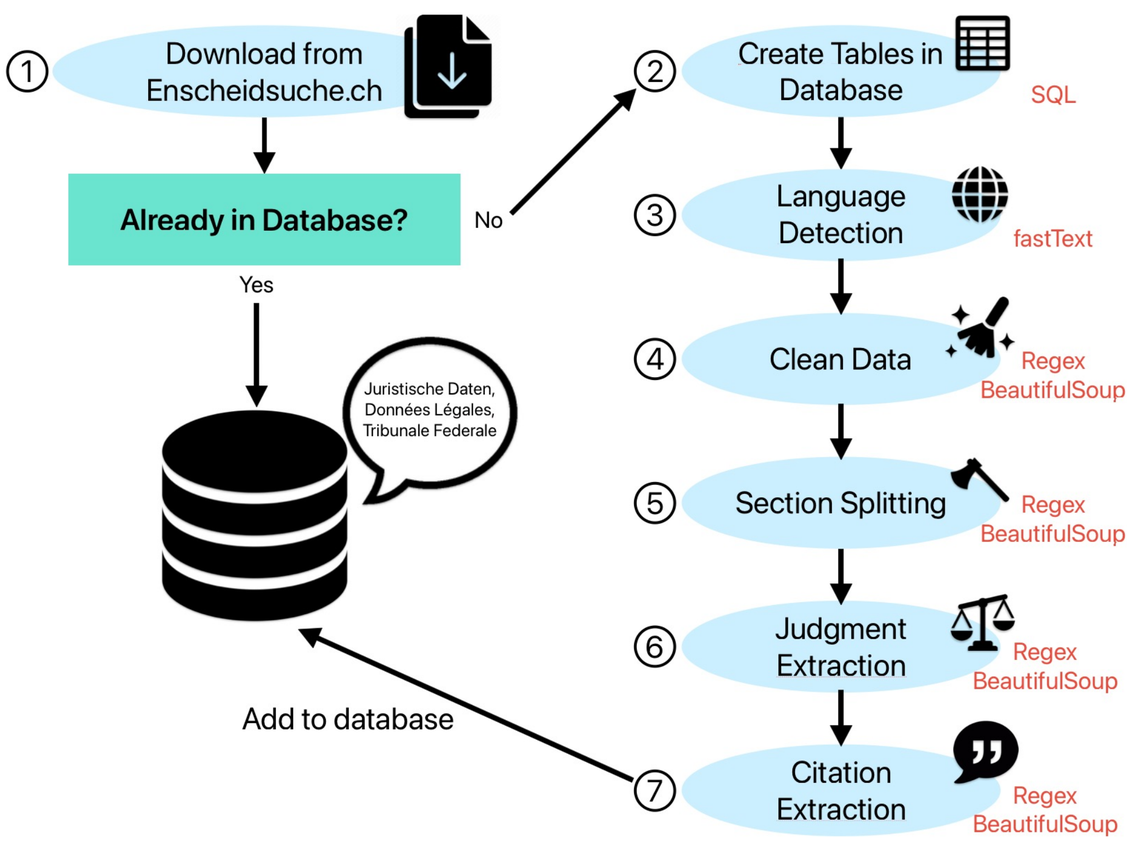
Example CVG German: GPT-3-Turbo Fine-Tuned

Das Bundesverwaltungsgericht zieht in Erwägung: 1. 1.1 Gemäss Art. 31 VGG beurteilt das Bundesverwaltungsgericht Beschwerden gegen Verfügungen nach Art. 5 VwVG. [...] 1.2 Beurteilt das Bundesverwaltungsgericht Beschwerden nach Art. 31 VGG, so wendet es unter Vorbehalt der Anwendung von staatsvertraglichen Rechten materielles Bundesrecht an (Art. 37 Abs. 1 VGG). 1.3 Mit Beschwerde kann die Verletzung von Bundesrecht (einschliesslich die unrichtige oder unvollständige Feststellung des rechtserheblichen Sachverhalts [Art. 97 Abs. 1 BGG] und die Unangemessenheit [Art. 105 Abs. 2 BGG]), die Rechtsverletzung nach Art. 95 BGG und die unrichtige oder unzweckmässige Sachverhaltsfeststellung (Art. 97 Abs. 1 BGG) gerügt werden. [...]

Translated to English

The Federal Administrative Court considers: 1. 1.1 In accordance with Art. 31 VGG, the Federal Administrative Court assesses appeals against decisions under Art. 5 VwVG. [...] 1.2 If the Federal Administrative Court assesses appeals in accordance with Art. 31 VGG, it shall apply substantive federal law, subject to the application of constitutional rights (Art. 37 para. 1 VGG). 1.3 An appeal may be lodged for a violation of federal law (including incorrect or incomplete determination of the legally relevant facts [Art. 97 para. 1 FSCA] and inappropriateness [Art. 105 para. 2 FSCA]), violation of the law pursuant to Art. 95 FSCA and incorrect or inappropriate determination of the facts (Art. 97 para. 1 FSCA). [...]

Results: Models replicate output style well, but fail to produce coherent, logical legal reasoning.

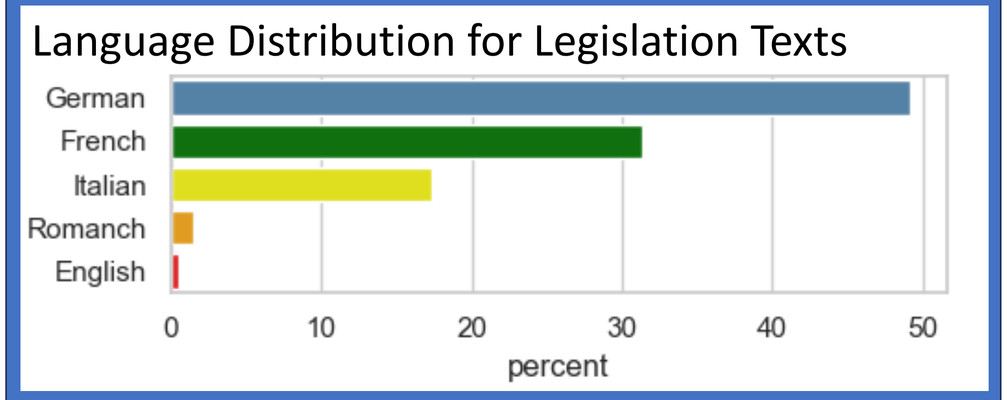


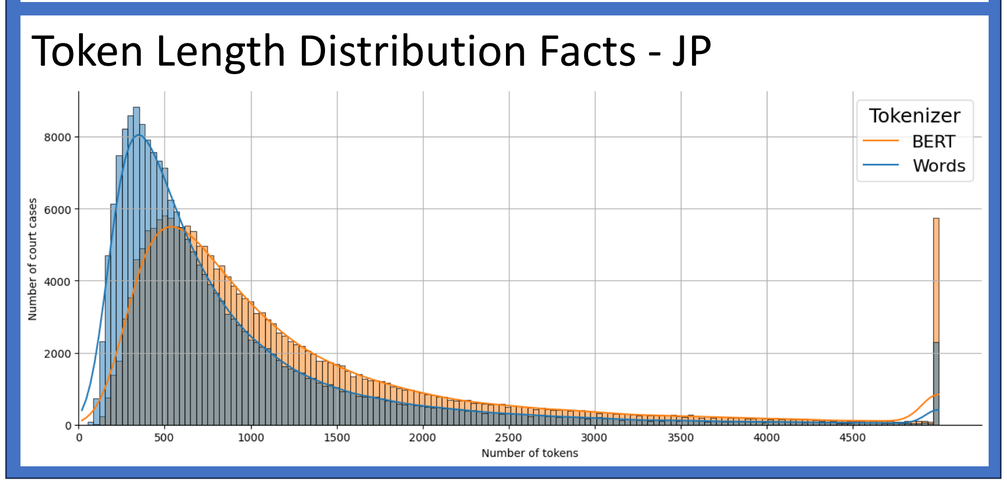
Paper

Models

Datasets

Dataset Properties Examples:





Conclusion

- End-to-end benchmark
- Challenge models on five key aspects
- Models, including ChatGPT, show low performance
- Results highlight opportunities for improving models and set the stage for next-generation LLM evaluations in domain-specific, multilingual contexts.