SwiLTra-Bench: The Swiss Legal Translation Benchmark

Joel Niklaus[1], Jakob Merane[2], **Luka Nenadic[2]**, **Sina Ahmadi[3]**, Yingqiang Gao[3], Cyrill A. H. Chevalley[4], Claude Humbel[3], Christophe Gösken[2], Lorenzo Tanzi[5], Thomas Lüthi[6], Stefan Palombo[1], Spencer Poff[1], Boling Yang[1], Nan Wu[1], Matthew Guillod[1], Robin Mamié[7], Daniel Brunner[7], Julio Pereyra[1], Niko Grupen[1]

[1]Harvey; [2]ETH Zurich; [3]University of Zurich; [4]University of Basel; [5]University of Geneva; [6]Canton of Solothurn; [7]Swiss Federal Supreme Court

1 Introduction

- **Four official languages** in Switzerland
- Legal translation is a **complex task**
- (Partially) automating legal translations improves **efficiency and access to justice**

2 Research Questions

1. What models are **best at translating** Swiss laws, case summaries, and press releases?
2. Can frontier model performance be reached by **fine-tuning** smaller models?

3 Dataset

Over **180K aligned Swiss legal translation pairs** (laws, case summaries, and press releases)

(a) CH-Law-Trans dataset.

Source	Split	#file	#de	#fr	#it	#rm	#en
Law	Train	5,206	5,206	5,206	5,206	51	219
	Valid	10	10	10	10	10	10
	Test	20	20	20	20	20	20
Article	Train	129,070	126,308	127,049	126,223	8,680	16,347
	Valid	789	785	785	784	785	785
	Test	740	738	738	738	738	738
Paragraph	Train	153,970	145,106	146,953	145,267	19,556	32,499
	Valid	1,490	1,441	1,438	1,437	1,441	1,439
	Test	1,214	1,176	1,178	1,178	1,177	1,176

Table 1: Overall corpus statistics for laws

4 SwiLTra-Judge

LLM-based method aligned with **expert annotations**

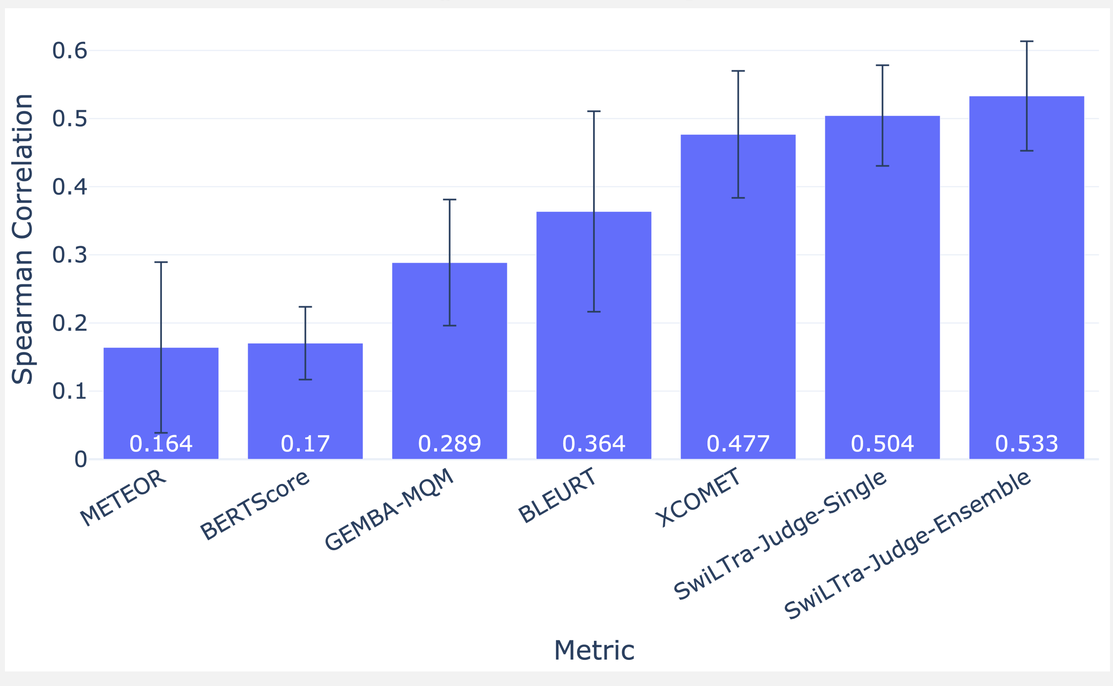

Figure 1: Spearman correlations with expert scores

5 Fine-Tuned Smaller Models

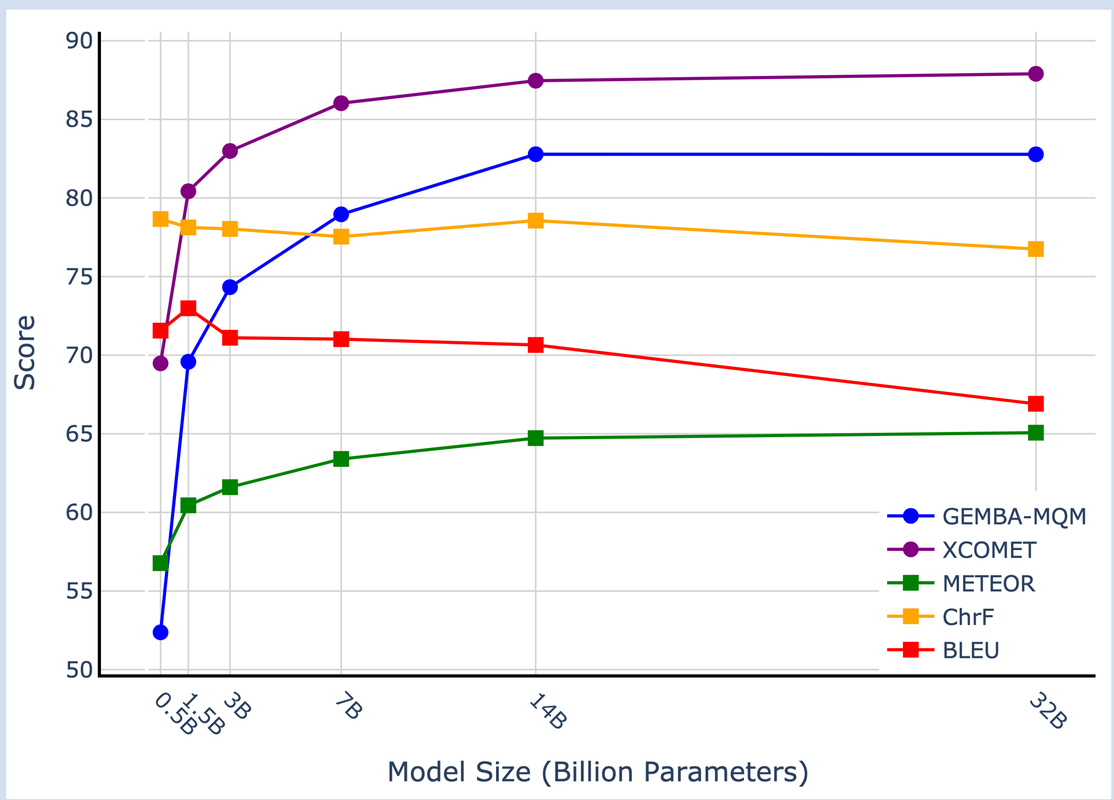

Figure 2: Metrics vs. model size for fine-tuned Qwen models

6 Main Results

- **No single model** wins in all tasks
- However, **frontier models** achieve superior translation performance across the board
- **Specialized translation models** excel specifically at translating laws (but less so at the other tasks)

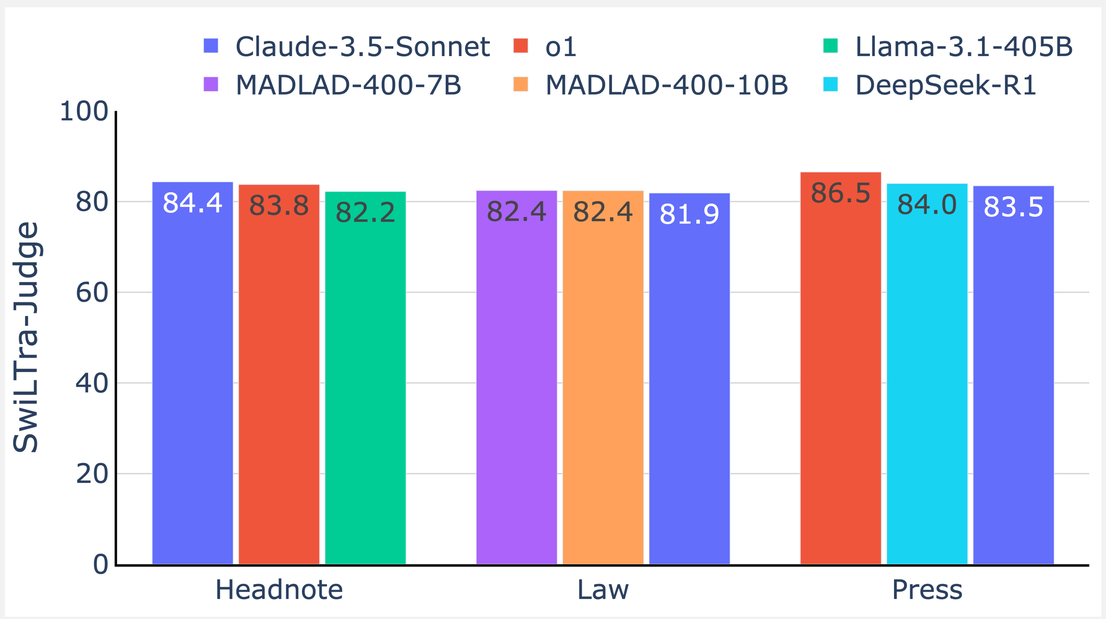

Figure 3: Best models per task

7 Summary

- **Complexity** of legal translations
- **SwiLTra-Bench**: Large-scale benchmark of over 180K aligned Swiss legal translation pairs
- Comprehensive **model comparisons**
- **SwiLTra-Judge**: LLM-based method based on expert annotations to automate evaluations