SwiLTra-Bench: The Swiss Legal Translation Benchmark

Joel Niklaus[Harvey]

Jakob Merane[ETH, Unil, MPI] **Luka Nenadic**[ETH] **Sina Ahmadi**[UZH] **Yingqiang Gao**[UZH]

Cyrill A. H. Chevalley[Basel] **Claude Humbel**[UZH] **Christophe Gösken**[ETH] **Lorenzo Tanzi**[Geneva]

Thomas Lüthi[Solothurn] **Stefan Palombo**[Harvey] **Spencer Poff**[Harvey] **Boling Yang**[Harvey] **Nan Wu**[Harvey]

Matthew Guillod[Harvey] **Robin Mamié**[SFSC] **Daniel Brunner**[SFSC] **Julio Pereyra**[Harvey] **Niko Grupen**[Harvey]

[Harvey]Harvey [ETH]ETH Zurich [SFSC]Swiss Federal Supreme Court [UZH]University of Zurich

[Basel]University of Basel [Geneva]University of Geneva [Unil]University of Lausanne

[Solothurn]Canton of Solothurn [MPI]Max Planck Institute for Research on Collective Goods

Correspondence: joel@niklaus.ai

Abstract

In Switzerland legal translation is uniquely important due to the country's four official languages and requirements for multilingual legal documentation. However, this process traditionally relies on professionals who must be both legal experts and skilled translators—creating bottlenecks and impacting effective access to justice. To address this challenge, we introduce SwiLTra-Bench, a comprehensive multilingual benchmark of over 180K aligned Swiss legal translation pairs comprising laws, headnotes, and press releases across all Swiss languages along with English, designed to evaluate LLM-based translation systems. Our systematic evaluation reveals that frontier models achieve superior translation performance across all document types, while specialized translation systems excel specifically in laws but under-perform in headnotes. Through rigorous testing and human expert validation, we demonstrate that while fine-tuning open SLMs significantly improves their translation quality, they still lag behind the best zero-shot prompted frontier models such as Claude-3.5-Sonnet. Additionally, we present SwiLTra-Judge, a specialized LLM evaluation system that aligns best with human expert assessments.

🤗 **Datasets** **Code**

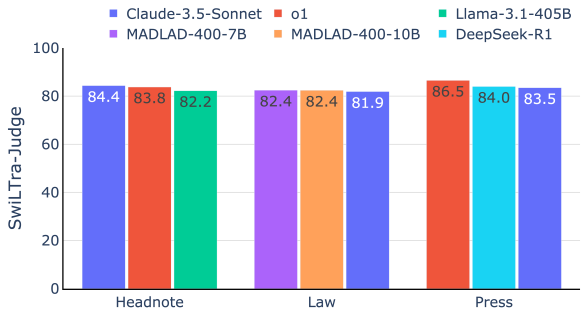


Figure 1: Best models per task.

1 Introduction

Neural Machine Translation (NMT) is one of the most studied Natural Language Processing (NLP) tasks. From encoder-decoder pipelines (Dai and Le, 2015; Vaswani et al., 2017) to modern decoder-only models (Brown et al., 2020; Touvron et al., 2023) NMT systems based on large language models (LLMs) have in recent years achieved notable advancements in translating texts across various genres (Ou et al., 2023; Zhang et al., 2024; Han et al., 2024) and in both high- and low-resource languages (Moslem et al., 2023; Vilar et al., 2023; Alves et al., 2023; Oliver et al., 2024). Nevertheless, the shortage of high-quality multilingual parallel legal translation data for training LLMs has hindered the performance of state-of-the-art NMT systems in translating legal texts. This limitation is primarily due to the discourse structures (Wiesmann, 2019) and specialized terminology (Katz et al., 2023) of legal texts, which consequently result in the current limited degree of automation for translation in the legal domain.

In multilingual countries like Switzerland, where legal documents are primarily translated manually by experts, developing reliable NMT systems for legal texts would significantly improve governmental efficiency and reduce administrative bottlenecks (Martínez-Domínguez et al., 2020). Beyond operational benefits, such systems could democratize access to legal information by enabling faster and more cost-effective translations across multiple national languages. Especially in lower-resourced languages like Romansh where full translation coverage is not currently economical, support from high-quality NMT systems could be game-changing. This broader accessibility would enhance the transparency of political decision-making and promote more inclusive civic participation (Moniz and Escartín, 2023). The potential impact extends beyond government operations to the private sector where law firms and businesses operating across linguistic regions could benefit from improved legal translation capabilities, potentially reducing costs and accelerating legal processes while maintaining accuracy. Although initial efforts have been made to develop NMT systems for translating Swiss legal documents (Martínez-Domínguez et al., 2020; Canavese and Cadwell, 2024), it remains unclear how well current LLMs perform on large benchmarks for translating Swiss legal texts, both in zero-shot and fine-tuning settings.

To address the shortage of Swiss legal training data and advance legal translation, we present three main contributions:

1. SWILTRA-BENCH: A large-scale benchmark of over 180K aligned Swiss legal translation pairs (laws, court decisions, press releases) spanning five languages (the four official Swiss languages plus English), substantially expanding available training data.
2. Comprehensive Model Comparison: The first large-scale evaluation of frontier LLMs and fine-tuned open SLMs on Swiss legal translations in both zero-shot and fine-tuning settings, providing insights into their relative strengths.
3. SWILTRA-JUDGE: An LLM-based method aligned with human expert annotations, offering a reliable automated framework to assess translation quality.

Our main findings are: a) frontier models consistently perform well across translation tasks; b) translation specific systems like MADLAD-400 are strong on laws but fall behind on headnotes; c) fine-tuning open SLMs drastically improves their translation quality but they are still behind zero-shot prompted frontier models; d) translation quality is uniform across languages; and e) agreement among human experts is higher for law translations than for headnotes.

2 SwiLTra-Bench

To support research in NMT, text alignment, and legal document processing, we present SwiLTra-Bench—a dataset uniting original legal texts and press releases on key court rulings.

(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

(b) CH-Headnote-Trans dataset.

Source	Split	#file	#de	#fr	#it
BGE	Train	13,330	13,330	13,330	13,330
	Valid	1,900	1,900	1,900	1,900
	Test	3,801	3,801	3,801	3,801
Regest	Train	13,550	13,550	13,550	13,550
	Valid	1,924	1,924	1,924	1,924
	Test	3,890	3,890	3,890	3,890
Text	Train	26,008	26,008	26,008	26,008
	Valid	3,805	3,805	3,805	3,805
	Test	7,316	7,316	7,316	7,316

(c) CH-Press-Trans dataset.

Source	Split	#file	#de	#fr	#it
Press	Train	867	867	867	152
	Valid	100	100	100	100
	Test	200	200	200	200

Table 1: Overall SWILTRA-BENCH corpus statistics. #file indicates the total number of files collected, while #de, #fr, #it, #rm, and #en represent the ones in the respective languages.

2.1 Data Collection

SwiLTra-Bench contains three sub-datasets:

1. Swiss Law Translations (CH-Law-Trans), including law-level (entire legal documents), article-level (individual articles), and paragraph-level (paragraphs within articles) translations.
2. Headnote Translations (CH-Headnote-Trans) of Swiss Supreme Court landmark court decisions (*"Bundesgerichtsentscheide" (BGE)* in German, *"Arrêts du Tribunal fédéral" (ATF)* in French, and *"Decisioni principali del Tribunale federale svizzero" (DFT)* in Italian) at the BGE-level (complete summaries of court decisions), regest-level (summaries focused on core legal issues), and text-level (detailed extraction of specific legal statements).
3. Swiss Supreme Court Press Release Translations (CH-Press-Trans).

All datasets contain parallel translations in German (de), French (fr), and Italian (it). Additionally, for CH-Law-Trans, some documents contain translations in Romansh (rm) and English (en).

We provide details of the data structure with concrete dataset examples in Appendix D.

2.2 Dataset Splits

We first segment each dataset by a unique identifier (entire laws, entire headnotes, and entire press releases) to ensure that no single law, headnote, or press release is split across training, validation, and test. For laws, we prioritize examples for the validation and test splits that (1) have more language versions (to guarantee good multilingual coverage), (2) have an official abbreviation (since abbreviations are only set for those laws that are presumed to be cited frequently[1], which we consider a good proxy for practical importance, (3) have shorter text lengths to make evaluation faster and cheaper, and (4) have newer applicability dates so that more recent and multilingual laws are prioritized for validation and testing, resulting in a more realistic evaluation setting. For headnotes, we similarly prioritize those with more recent publication years for validation and test. Finally, for press releases, we focus on maximizing multilingual coverage by ensuring all validation and test examples are available in all present languages (German, French and Italian). The training sets contain all examples not held out for validation or testing.

2.3 Data Statistics

Table 1 presents the overall statistics of the three datasets included in SwiLTRa-Bench. We visualize the training set text lengths for the shortest levels used for training and evaluation in Figure B.1. For completeness, we show histograms for all levels in Figure B.2 and Figure B.3. To calculate these statistics, we used an NLTK[2] word tokenizer, splitting sentences based on whitespace and punctuation.

Existing parallel legal corpora use automated methods for sentence alignment (Koehn, 2005; Ziemski et al., 2016). In SwiLTRa-Bench, we rely on the structure provided by the official government bodies such as law paragraphs embedded in the HTML, resulting in high-quality alignment.

3 Experimental Setup

3.1 Evaluation

To paint a representative picture of translation capabilities, we evaluate models across five main classes: 1) translation models, i.e., models specifically trained for translation tasks, 2) frontier models, i.e., large foundation models pre-trained on web-scale data and post-trained on diverse tasks, 3) reasoning models, i.e., models using significant resources at test time to improve output quality, 4) open models, i.e., typically small language models (SLMs) with publicly available weights, and 5) fine-tuned models, i.e., models specifically fine-tuned on SwiLTRa-Bench. We conducted our evaluation using the `lighteval` framework due to its ease of use and good support for custom metrics.[3]

3.1.1 Metrics

We evaluated translations using lexical (BLEU (Papineni et al., 2002), ChrF (Popović, 2015), METEOR (Banerjee and Lavie, 2005)) and model-based metrics (BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), XCOMET (Guerreiro et al., 2024), GEMBA-MQM (Kocmi and Federmann, 2023)). Due to the 512-token limit, BLEURT and XCOMET cannot process press releases. Given GEMBA-MQM's strong correlation with human judgments, we prioritized it alongside XCOMET, METEOR, and ChrF, ensuring both lexical and trained metrics for diversity.

3.2 Fine-tuning

To provide an overview of the current open SLM landscape, we fine-tuned Gemma-2 2B and 9B (Team et al., 2024), Llama 1B, 3B and 8B (Grattafiori et al., 2024), Phi-3.5 mini and Phi-3 medium (Abdin et al., 2024), and Qwen2.5 0.5B, 1.5B, 3B, 7B, 14B and 32B (Team, 2024) models on our dataset. We fine-tuned using Hugging Face `transformers`[4] and `unsloth`[5] using 4-bit quantization and 8bit AdamW (Loshchilov and Hutter, 2019; Dettmers et al., 2022) on a single 80GB NVIDIA H100 GPU. We used rank stabilized LoRA (Hu et al., 2021; Kalajdzievski, 2023) with rank 16 and alpha 16. We trained with the model's native chat template on sequence length 512, covering more than 99% of the training dataset and truncating the rest. The instruction template was simply:

{source language}: {source text}
{target language}: {target text}

We trained on the entire training set for maximal coverage of the data. We used packing, weight

[1] https://www.bk.admin.ch/apps/gtr/de/index.html
[2] https://www.nltk.org
[3] https://github.com/huggingface/lighteval
[4] https://github.com/huggingface/transformers
[5] https://github.com/unslothai/unsloth

decay 0.01, batch size 128 and early stopping with patience 3. In most cases, the lowest evaluation loss is reached after exactly 1 epoch. We used a linear learning rate schedule with 1000 warmup steps and learning rate $1e-4$. We manually tuned the learning rate ($1e-5$ - $1e-3$), weight decay (0.01, 0.1), label smoothing (factor 0, 0.01, 0.1) and LoRA rank (16, 128). We used the train and validation sets of the Law and Headnote translations on the lowest (shortest) levels, i.e., the paragraph and text levels. For all fine-tuned models, we used the instruction-tuned variant since they have shown to better adapt to new tasks (Niklaus et al., 2024).

4 Results and Analysis

In the tables, we bolded the highest and underlined the second highest score per metric. Unless stated otherwise, we excluded Romansh from the evaluations to ensure comparability, since it is not supported by the translation models. Unless stated otherwise, results are averaged over source languages, target languages, and tasks. In general, we considered the law and headnote translation tasks at the highest granularity (paragraph-level and text-level) so we can compare all model categories (translation models and fine-tuned models are optimized for shorter sequence lengths). We show all metrics with standard errors obtained through bootstrapping. Higher values are better for all metrics.

4.1 Statistical Significance Testing

The model comparisons presented throughout this section are descriptive in nature. When we state that one model "outperforms" or "outcompetes" another, we are describing observed differences in the reported metrics; we do not claim statistical significance unless explicitly stated. We only apply null hypothesis significance testing (NHST) in those instances where we explicitly report statistically significant results. In such cases, we provide information about the tests used, statistics obtained, and effect sizes.

4.2 Translation Models

We compare translation models in Table 2. Surprisingly, Google-Translate performs poorly compared to open translation models like MADLAD-400 (Kudugunta et al., 2024) and Tower-Instruct. Facebook's SeamlessM4T (Communication et al., 2023) model's text-to-text capabilities also underwhelm. MADLAD-400 performs very well, outperforming GPT-4o on XCOMET. The Tower (Alves et al., 2024) models land somewhere in between.

Model	Size	↑ GEMBA-MQM	↑ XCOMET	↑ METEOR	↑ ChrF
Google-Translate	N/A	53.20 ± 0.2	64.61 ± 0.1	41.15 ± 0.1	47.81 ± 0.1
MADLAD-400-3B	3B	**62.89 ± 0.1**	<u>86.82 ± 0.1</u>	42.44 ± 0.1	51.36 ± 0.1
MADLAD-400-7B	7B	<u>62.66 ± 0.1</u>	**87.40 ± 0.1**	<u>43.70 ± 0.1</u>	<u>51.67 ± 0.1</u>
MADLAD-400-10B	10B	61.46 ± 0.1	86.65 ± 0.1	43.10 ± 0.1	**52.24 ± 0.1**
SeamlessM4T	2B	23.35 ± 0.2	43.03 ± 0.1	37.81 ± 0.1	24.90 ± 0.1
TowerInstruct-7B	7B	54.04 ± 0.2	72.97 ± 0.1	41.65 ± 0.2	43.00 ± 0.1
TowerInstruct-13B	13B	57.38 ± 0.2	75.94 ± 0.1	**43.95 ± 0.2**	48.46 ± 0.1

Table 2: Translation models across different families and sizes.

4.3 Frontier Models

We show results for frontier and reasoning models in Table 3. GPT-4o underperforms both of its peers Claude-3.5-Sonnet and Llama-3.1-405B. This is particularly unexpected, as models tend to favor their own completions (Panickssery et al., 2024), and GEMBA-MQM is operated by GPT-4o. Claude-3.5-Sonnet demonstrates strong performance, competing closely with o1, the top-performing model. Surprisingly, o1-mini performs only on par with the other models at the smaller scale and even underperforms Claude-3.5-Haiku. Overall, Anthropic's models are really strong, and even more so from a cost-to-performance perspective compared to reasoning models like o1.

4.4 Fine-tuned Models

Fine-tuning leads to notable performance gains (see Appendix Table C.1). Figure 2 presents fine-tuned models' performance across various sizes. The two Gemma models and particularly the Llama 1B and 3B models, advance the Pareto frontier, though performance starts to flatten at the 3B scale and plateaus after 9B parameters. Interestingly, both Phi models clearly underperform their peers.

4.5 Performance Progression by Model Size

The Qwen2.5 model family, with six sizes from 0.5B to 32B parameters, is ideal for studying perfor-

Model	Size	↑ GEMBA-MQM	↑ XCOMET	↑ METEOR	↑ ChrF
Claude-3.5-Sonnet	large	80.66 ± 0.2	<u>90.70 ± 0.1</u>	56.71 ± 0.2	65.87 ± 0.1
DeepSeek-V3	large	80.04 ± 0.2	89.77 ± 0.1	56.60 ± 0.1	69.99 ± 0.1
DeepSeek-R1	large	77.90 ± 0.2	84.36 ± 0.1	55.79 ± 0.1	69.12 ± 0.1
GPT-4o	large	80.27 ± 0.2	80.96 ± 0.1	55.56 ± 0.1	63.27 ± 0.1
Gemini-1.5-Pro	large	81.88 ± 0.2	87.13 ± 0.1	<u>57.92 ± 0.1</u>	<u>70.07 ± 0.1</u>
Llama-3.1-405B	large	81.59 ± 0.1	89.37 ± 0.1	54.48 ± 0.1	68.07 ± 0.1
Mistral-Large	large	81.88 ± 0.2	87.04 ± 0.1	54.86 ± 0.1	63.71 ± 0.1
o1	large	**85.81 ± 0.1**	**91.35 ± 0.1**	**58.91 ± 0.1**	**70.11 ± 0.1**
Claude-3.5-Haiku	small	80.40 ± 0.2	88.84 ± 0.1	52.15 ± 0.2	61.09 ± 0.1
GPT-4o-mini	small	<u>82.59 ± 0.2</u>	87.90 ± 0.1	54.03 ± 0.1	59.86 ± 0.1
Gemini-1.5-Flash	small	80.76 ± 0.2	85.33 ± 0.1	55.35 ± 0.1	65.44 ± 0.1
Llama-3.3-70B	small	79.25 ± 0.2	88.02 ± 0.1	53.43 ± 0.1	65.92 ± 0.1
Mistral-Small	small	81.69 ± 0.2	87.04 ± 0.1	54.83 ± 0.1	63.66 ± 0.1
o1-mini	small	81.96 ± 0.2	87.46 ± 0.1	53.34 ± 0.1	59.32 ± 0.1

Table 3: Frontier models across different families and sizes.

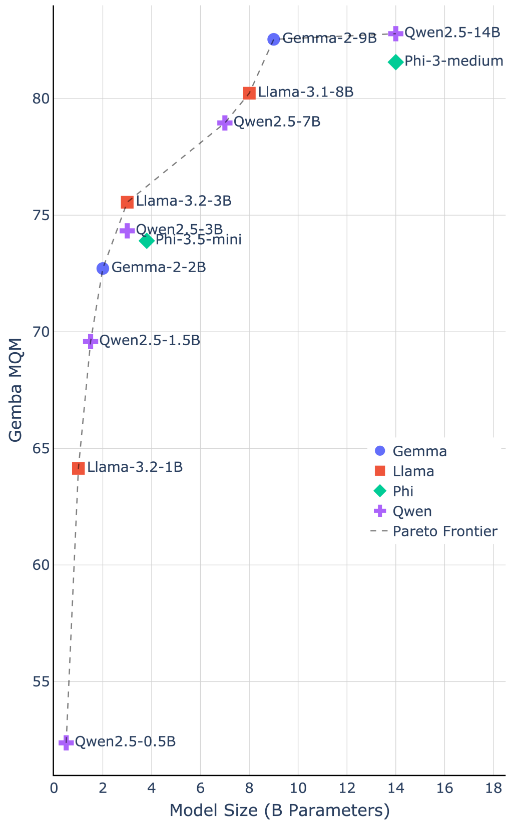


Figure 2: Finetuned models across different sizes.

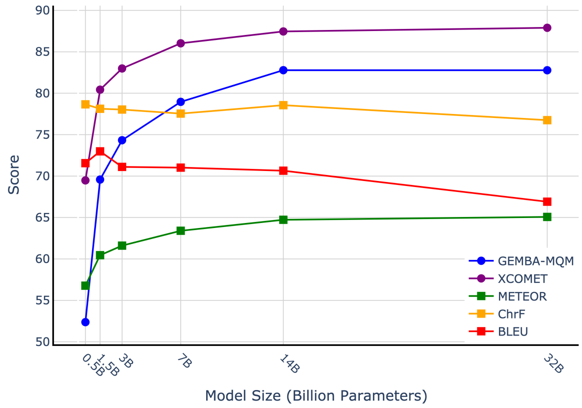


Figure 3: Lexical (square) and model-based (circle) metrics vs model size for finetuned Qwen models.

Model	Category	Task	↑ GEMBA-MQM	↑ METEOR	↑ ChrF
o1	reasoning	Headnote	**93.50 ± 0.1**	60.89 ± 0.1	62.62 ± 0.2
o1	reasoning	Law	<u>91.11 ± 0.1</u>	55.87 ± 0.1	66.84 ± 0.1
o1	reasoning	Press	64.62 ± 0.4	59.28 ± 0.3	**78.38 ± 0.1**
Claude-3.5-Sonnet	frontier	Headnote	88.65 ± 0.1	61.39 ± 0.1	63.96 ± 0.2
Claude-3.5-Sonnet	frontier	Law	85.71 ± 0.1	52.16 ± 0.1	73.15 ± 0.1
Claude-3.5-Sonnet	frontier	Press	60.83 ± 0.8	55.29 ± 0.5	55.47 ± 0.1
MADLAD-400-7B	translation	Headnote	80.54 ± 0.2	57.71 ± 0.1	67.49 ± 0.2
MADLAD-400-7B	translation	Law	85.06 ± 0.2	57.09 ± 0.2	61.86 ± 0.2
SLT-Qwen2.5-32B	finetuned	Headnote	82.58 ± 0.1	**66.56 ± 0.1**	75.17 ± 0.2
SLT-Qwen2.5-32B	finetuned	Law	80.80 ± 0.2	<u>64.41 ± 0.1</u>	<u>76.90 ± 0.1</u>
Qwen2.5-14B	open	Headnote	69.88 ± 0.2	47.09 ± 0.1	53.58 ± 0.2
Qwen2.5-14B	open	Law	63.04 ± 0.2	34.33 ± 0.1	52.02 ± 0.1

Table 4: Best models per category across different tasks.

mance progression over model size. We analyzed fine-tuned Qwen models up to 32B using five metrics (two model-based, three lexical) in Figure 3. METEOR is the only lexical metric well correlated with XCOMET and GEMBA-MQM. All three confirm a clear trend that larger models produce higher-quality translations. GEMBA-MQM shows the largest score range (GEMBA-MQM: 52.4 - 82.8 vs XCOMET: 69.5 - 87.9 and METEOR: 56.8 - 65.1) and making it most useful for differentiating models. Interestingly, both ChrF and BLEU are negatively correlated with the model-based metrics on this task for the fine-tuned Qwen models. Beyond the inherent subjectivity in assessing translation quality, this may hint at the greater importance of a legal text's conveyed meaning over the mere use of certain exact terms.

4.6 Comparison Across Tasks

In Table 4, we show the best models' performance per category across tasks. The best open small model falls far behind the others but, with fine-tuning, overtakes the best translation model. It matches the smaller frontier models but still lags behind the larger ones. All models except MADLAD-400-7B perform better on headnote than law translation. While Sonnet competes with o1 on headnote and law translation, it falls off on press releases.

4.7 Comparison Across Languages

In Figure 4, we compare the best models per category across language directions on CH-Law-Trans. Performance to and from German, French, and Italian is homogeneous across models. When translating from English to the other languages, all models perform worse than from the three main Swiss languages. Since the English source texts are already translations and are not legally binding, the federal translators may have applied less rigor in generating them, potentially resulting in lower quality and slight deviations. Anecdotally, the lawyers co-authoring this work confirm that the English source texts are occasionally less precise. So, the lower scores may also indicate that the judge model bases its grading on imperfect source text.

Romansh is a low-resource language and only spoken by less than 50K people in Switzerland.[6] As our dataset consists of the entire data readily available for federal laws, Supreme Court headnotes, and Supreme Court press releases in Switzer-

[6] https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/sprachen-religionen/sprachen.html

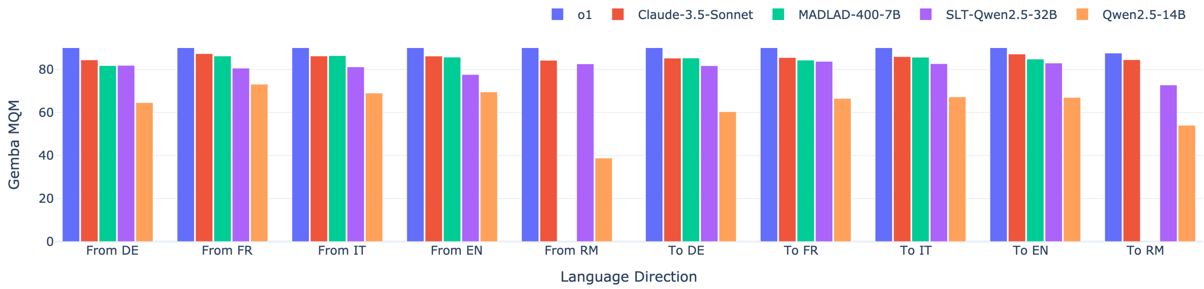


Figure 4: Best models per category across languages.

land, it is not possible to extend the Romansh coverage. Adding additional data sources is not a viable option due to likely lower data quality and limited sentence level alignment across languages. Romansh is not supported by most translation models such as MADLAD-400. In our opinion, the fact that these models don't support Romansh at the moment highlights the value of our dataset for low-resource languages. Indeed, our dataset now enables teams building translation models to train and evaluate on >160K and >8K Romansh translation pairs respectively. Surprisingly, o1 and Sonnet still perform very well when translating from Romansh to other languages. When translating to Romansh, all models' quality drops off, sometimes sharply. Perhaps similarly to humans, also for LLMs speaking or writing a language seems harder than understanding it.

5 Expert Evaluation

To study how well human legal experts agree with the automated metrics, we conducted an expert evaluation. All experts are authors of the paper; the majority are doctoral candidates, and all hold at least a Bachelor's degree in Swiss law. We only evaluated the laws and headnotes since they are much shorter and we could evaluate more examples in the time available. Due to limited expert time, we selected the top model from four categories: frontier (Claude-3.5-Sonnet), reasoning (o1), translation (MADLAD-400-7B) and fine-tuned (SLT-Qwen2.5-32B).

The experts were asked to assign a score between 0 and 100 to each translation. For this purpose, the experts were given a source text, its "gold translation" (official translation of the Swiss authorities) as a reference and a predicted translation. The scores only reflected the completeness and accuracy of the predicted translation, with less emphasis on readability and other stylistic attributes. To ensure consistency, the experts agreed on a point deduction system in advance and discussed certain borderline cases (annotation guidelines are in Appendix G). In total, 200 translations were assigned a score by exactly two experts. Each expert assigned scores independently, without consulting the other annotators. For the expert agreement with judge metrics (see section 6) and for the evaluation of the best models (see subsection 5.2), we averaged the scores of the two annotators.

5.1 Inter-Annotator Agreement

The average Krippendorff's α was 0.56 for laws and 0.41 for headnotes. Agreement was generally higher for laws than for headnotes across most language pairs and tasks. This consistent pattern is unlikely due to a calibration issue and instead suggests that the headnote task inherently allows for more subjectivity. This is likely due to headnotes being longer on average and containing greater interpretive nuance, which can lead annotators to miss or accord different importance to different details. We also hypothesize that laws inherently exhibit a more structured and clearer pattern, making them easier to translate and evaluate.

The moderate inter-annotator agreement suggests that, despite clear instructions, a certain degree of subjectivity was inherent in the task. In addition, we observed smaller differences between the individual language pairs, suggesting that not all annotators were perfectly aligned. However, disagreements tended to be minor and were rarely fundamental. In Figure 5 we show the absolute point difference between the two annotators evaluating the same samples. In almost half of the cases the two annotators completely agree and in 92% the difference is smaller than 30 points.

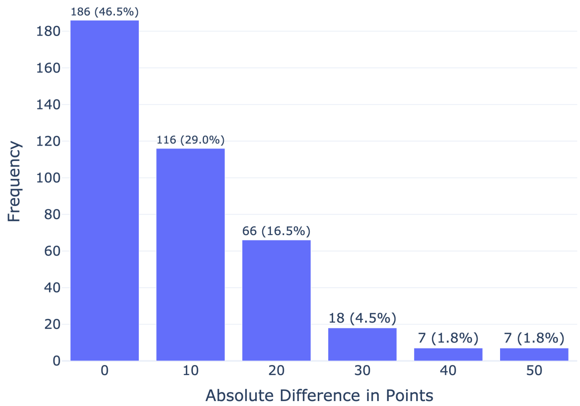


Figure 5: Absolute point difference between annotators.

5.2 Which Model is the Best?

In Table 5 we show the expert scores together with the best metrics for the best models per category. It is evident here that XCOMET aligns best with the experts. We conclude that both for translating laws and headnotes Claude 3.5 Sonnet is the best model followed by o1 for laws and both o1 and the finetuned Qwen2.5-32B model for headnotes.

Model	Task	↑ Experts	↑ XCOMET	↑ BLEURT	↑ GEMBA-MQM
Claude-3.5-Sonnet	Headnote	89.21 ± 2.2	90.91 ± 1.5	28.96 ± 3.7	86.53 ± 1.6
Claude-3.5-Sonnet	Law	**94.55 ± 1.1**	**93.30 ± 1.1**	<u>34.16 ± 3.2</u>	88.86 ± 1.2
MADLAD-400-7B	Headnote	71.77 ± 3.3	85.57 ± 2.8	12.20 ± 3.1	76.13 ± 4.9
MADLAD-400-7B	Law	83.77 ± 2.8	89.42 ± 2.3	28.97 ± 3.5	88.63 ± 2.0
SLT-Qwen2.5-32B	Headnote	84.86 ± 2.4	88.62 ± 2.1	30.78 ± 4.3	75.89 ± 4.2
SLT-Qwen2.5-32B	Law	85.74 ± 2.1	88.03 ± 2.2	**36.42 ± 3.9**	81.78 ± 2.5
o1	Headnote	84.29 ± 2.1	89.58 ± 1.8	16.77 ± 4.2	<u>92.34 ± 1.4</u>
o1	Law	<u>89.91 ± 1.5</u>	<u>92.33 ± 1.5</u>	28.19 ± 3.2	**92.97 ± 1.0**

Table 5: Expert scores for best models across categories

6 SwiLTra-Judge

Automatic evaluation of natural language generation is challenging. Lexical metrics like BLEU or METEOR correlate weakly with human judgments (Zhang et al., 2020). Early model-based metrics such as BERTScore or BLEURT perform better, but recently, LLM-as-Judge has emerged as the dominant paradigm (Zheng et al., 2023). Each task, however, is unique and requires its own judge setup. In this section, we ablate key aspects of the judge setup, including the judge model, prompt, and few-shot sample selection.

6.1 Setup

We use GPT-4o, GPT-4o-mini, Gemini-1.5-pro, and Gemini-1.5-flash in our judge model ablation. We also tested Claude Sonnet and Haiku as judges, but they failed to follow grading instructions.[7] The o1 and o1-mini models showed very low or even negative correlations with human judgments and are thus excluded. We randomly selected one few-shot example from the dev sets of laws, headnotes, and press releases. To ensure judge models saw diverse translation qualities, we chose models of varying strengths (Claude 3.5 Sonnet for laws, Mixtral-8x7B-Instruct-v0.1 for headnotes, and Qwen2.5-1.5B-Instruct for press releases). We used a simple prompt "*Translate to target-language*" to generate translations. Sample judgments per few-shot example were written by one lawyer author and double-checked by another. We tested two few-shot styles *single* (all examples in one language direction: fr-de) and *diverse* (law article en-it, headnote de-fr and press release fr-de). We ablated two user prompts with absolute grading (*basic* and *detailed*) and one with deduction grading similar to the codebook given to the human expert annotators (*codebook*). Judge prompts are in Appendix F.

We measured the correlation of our judge setups with the human expert scores on the 400 human annotated samples. To get a higher confidence signal, we removed samples where the two human experts disagreed by 30 points or more (32/400 or 8%). Find complete results in Table E.1. Unless specified otherwise, we report Spearman correlation with human judgments with cross validation. Based on our expert evaluation, we answer the following research questions (RQs):

RQ1: Are small models judges good enough?

A: Yes, the small models even outperform their larger counterparts. Over all tested configurations GPT-4o and GPT-4o-mini are tied at 0.41 ± 0.08 mean Spearman correlation. Gemini-1.5-flash even outperforms Gemini-1.5-pro as a judge model (0.33 ± 0.07 vs 0.27 ± 0.09). For the best configuration GPT-4o-mini even outperforms GPT-4o (0.48 ± 0.1 vs 0.45 ± 0.07) and the same holds for Gemini-1.5-flash vs Gemini-1.5-pro (0.5 ± 0.07 vs 0.3 ± 0.08).

RQ2: Is the deduction judgment style better than the absolute style?

A: Judges using the deduction style align more closely with human judgments. Across all configurations, there is little difference between the two absolute styles (0.33 ± 0.09 for *basic* and 0.32 ± 0.08

[7]They would insist on generating JSON output while we very clearly just asked for plain-text.

for *detailed* user prompt). However, the deduction style aligns much more closely with experts (0.42 ± 0.08). The top six highest correlating configurations all use the deduction style. This finding anecdotally confirms that LLM judge models reach judgments more similar to human experts when prompted in a more aligned way.

RQ3: Are few-shot examples in a single language pair sufficient, or is it necessary to include examples from diverse language pairs?

A: On average, the language directions of the few shot examples do not matter, but the best configuration uses diverse language directions. Across all 24 investigated configurations, there is only minimal difference between the single and diverse language direction setup (0.37 ± 0.08 vs 0.35 ± 0.08). However, the best configuration overall, uses diverse language directions.

RQ4: How does SwiLTra-Judge perform compared to other metrics?

A: Our SwiLTra-Judge exhibits the highest correlation with human judgments among tested translation metrics. Figure 6 shows Spearman correlation with human judgments for sample-level metrics (this excludes BLEU and ChrF). As expected, METEOR and BERTScore perform poorly, with correlations below 0.2. Surprisingly, the recent GEMBA-MQM metric both underperforms BLEURT and XCOMET. Our SwiLTra-Judge-Single (gemini-1-5-flash-codebook-diverse-deduction) is significantly better than the second-best metric XCOMET (0.5 ± 0.07 vs 0.48 ± 0.09, $p = 0.0008$). Our SwiLTra-Judge-Ensemble (a simple mean score from GPT-4o-mini and Gemini 1.5 Flash with codebook style and both single and diverse few-shot setup) is again significantly better than our SwiLTra-Judge-Single (0.53 ± 0.08 vs 0.5 ± 0.07, $p < 0.0001$). To compute both p-values, we have performed an unpaired t-test with $N = 368$ each.

6.2 Judge Harshness

In Figure 7 we show the average score over the best four generator models per category across judge models, system and few shot styles. We confirm here that the language directions of the few shot examples only has a minor effect. We see that the detailed system style leads to the harshest scores across models. Interestingly, Gemini-1.5-pro and GPT-4o-mini judge very similarly in terms of harshness. All models except GPT-4o judge more leniently with the codebook system style.

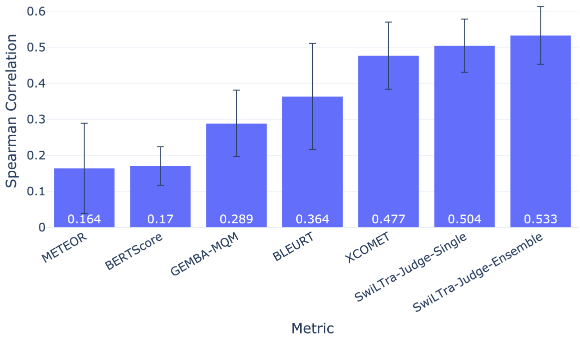


Figure 6: Spearman correlations with human expert scores.

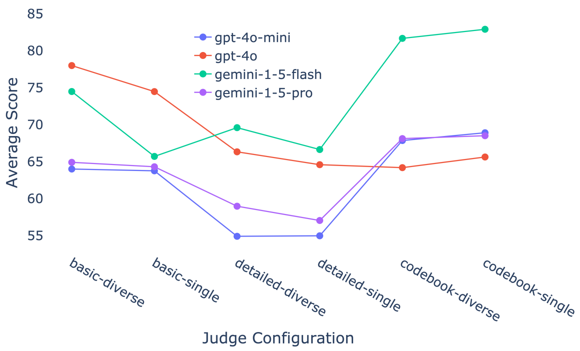


Figure 7: Judge harshness across configurations.

6.3 Best Model Per Task

Now that we built a trusted metric for our translation benchmark, we ran it over the entire dataset for all models. With this, we can recommend the best model for each task. In Figure 1 we show the top three models per task using SwiLTra-Judge as a metric. There are no large differences among top models, but the highest scores are achieved by Sonnet on headnotes, MADLAD-400-7B on laws and o1 on press releases. Sonnet ranks in the top 3 for all tasks. One reason for Sonnet's high scores could be that the few-shot example in the judge prompt with the highest score was translated by Sonnet, possibly making the judge models prefer its style. However, human experts clearly favored Sonnet without bias from few-shot examples.

7 Related Work

The application of NLP to legal texts has seen significant growth in recent years. This increased attention is driven by the growing need to automate and enhance legal processes, improve access to justice, and handle the vast amounts of legal documentation produced globally.

Recent research has explored various aspects of legal text processing. Legal judgment prediction has emerged as a crucial area, with studies demonstrating success across different jurisdictions, including the US (Semo et al., 2022), Europe (Vaudaux et al., 2023) and Switzerland (Niklaus et al., 2021, 2022). Notable advances have been made in verdict prediction (Medvedeva et al., 2020), topic classification (Papaloukas et al., 2021; Benedetto et al., 2023; Rasiah et al., 2023), and legal QA systems (Zhong et al., 2020). LegalBench (Guha et al., 2023) LexGLUE (Chalkidis et al., 2022) and LEXTREME (Niklaus et al., 2023) are established as comprehensive benchmark suites comprising multiple legal NLP tasks, including text classification, named entity recognition, and legal entailment across various jurisdictions and legal areas.

The translation of legal texts has significant societal impact and is increasingly important for training translators and practical applications, especially as machine translation gains prominence (Killman, 2024). However, legal translation poses challenges due to domain-specific terminology, reliability in legal formulae, and non-compliance with legal conventions (Killman, 2023; Giampieri, 2023). While some U.S. courts have considered NMT, it remains far from replacing human translators (Vieira et al., 2021). Robust NMT systems are essential for judicial and governmental services, with recent advancements leveraging pretrained LLMs and fine-tuning techniques (Zhu et al., 2024). Prior research has focused on legal NMT for languages like Chinese, and Arabic (Ding, 2024; ElFqih and Monti, 2023). However, a significant research gap remains in translating legal texts between Switzerland's national languages, which our work aims to address.

8 Conclusions

In this work, we introduced SwiLTra-Bench, a high-quality multilingual legal translation benchmark, and evaluated mainstream LLM-based NMT systems under both zero-shot and fine-tuned settings. Our analysis, validated by human expert annotations, showed that frontier models outperform all others, while translation-specific systems like MADLAD-400 excel on laws but struggle with headnotes. Fine-tuning open LLMs significantly improves their performance, though they still lag behind zero-shot frontier models, and translation quality remains consistent across Swiss languages. Finally, our SwiLTra-Judge model, optimized for legal translation evaluation, achieves the highest alignment with human expert judgments, providing a valuable automated metric for future research.

Limitations & Future Work

Our fine-tuned models are much stronger than the initial instruction-tuned open models they are based on, but they still under-perform large closed models. Future work could investigate techniques such as model merging (Yang et al., 2024) to further improve and bring them closer to the frontier models. While we evaluated a large variety of models, we could not evaluate them all. Future work could investigate other promising models such as Grok.[8] We took great care to validate our results with human expert studies. However, our resources were limited and we could not investigate certain languages (e.g., Romansh) and our sample sizes were still rather small. Future work could perform a more broad and in-depth human evaluation.

Ethics Statement

Our benchmark contains no personal, sensitive, or private information; it consists solely of publicly available data.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. Luka Nenadic's research is funded by SNSF grant No. 10002634.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, and *et al.* 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and André FT Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

[8] https://x.ai/blog/grok-2

Satanjeev Banerjee and Alon Lavie. 2005. 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. Association for Computational Linguistics.

Irene Benedetto, Gianpiero Sportelli, Sara Bertoldo, Francesco Tarasconi, Luca Cagliero, and Giuseppe Giacalone. 2023. On the use of pretrained language models for legal Italian document classification. In *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 27th International Conference KES-2023, Athens, Greece, 6-8 September 2023*, volume 225 of *Procedia Computer Science*, pages 2244–2253. Elsevier.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Paolo Canavese and Patrick Cadwell. 2024. Translators' perspectives on machine translation uses and impacts in the Swiss Confederation: Navigating technological change in an institutional setting. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 347–359.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4310–4330. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lijie Ding. 2024. A Comparative Study on the Quality of English-Chinese Translation of Legal Texts Between ChatGPT and Neural Machine Translation Systems. *Theory and Practice in Language Studies*, 14(9):2823–2833.

Khadija Ait ElFqih and Johanna Monti. 2023. On the Evaluation of Terminology Translation Errors in NMT and PB-SMT In the Legal Domain: A Study on the Translation of Arabic Legal Documents into English and French. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 26–35.

Patrizia Giampieri. 2023. Is machine translation reliable in the legal field? a corpus-based critical comparative analysis for teaching ESP at tertiary level. *Esp Today-Journal of English for Specific Purposes at Tertiary Level*, 11(1):119–137.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and *et al.* 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2404.14219*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit

Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2308.11462 [cs].

Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, 6:1211564.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. ArXiv:2106.09685 [cs].

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *CoRR*, abs/2312.03732.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Jeffrey Killman. 2023. Machine translation and legal terminology: Data-driven approaches to contextual accuracy. *Handbook of Terminology*, pages 485–510.

Jeffrey Killman. 2024. Machine translation literacy in the legal translation context: a SWOT analysis perspective. *The Interpreter and Translator Trainer*, 18(2):271–289.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: papers*, pages 79–86.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Rubén Martínez-Domínguez, Matīss Rikters, Artūrs Vasiļevskis, Mārcis Pinnis, and Paula Reichenberg. 2020. Customized neural machine translation systems for the Swiss legal domain. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 217–223.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the European Court of Human Rights. *Artif. Intell. Law*, 28(2):237–266.

Helena Moniz and Carla Parra Escartín. 2023. Towards responsible machine translation. *Ethical and Legal Considerations in Machine Translation. Cham: Springer*.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021, NLLP@EMNLP 2021, Punta Cana, Dominican Republic, November 10, 2021*, pages 19–35. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Joel Niklaus, Lucia Zheng, Arya D. McCarthy, Christopher Hahn, Brian M. Rosen, Peter Henderson, Daniel E. Ho, Garrett Honke, Percy Liang, and Christopher Manning. 2024. FLawN-T5: An Empirical Examination of Effective Instruction-Tuning Data Mixtures for Legal Reasoning. *arXiv preprint*. ArXiv:2404.02127 [cs].

Antoni Oliver, Sergi Alvarez-Vidal, Egon Stemle, and Elena Chiocchetti. 2024. Training an NMT system for legal texts of a low-resource language variety South Tyrolean German-Italian. In *Proceedings of* the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 573–579.

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on Greek legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2021, NLLP@EMNLP 2021, Punta Cana, Dominican Republic, November 10, 2021*, pages 63–75. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics. Event-place: Philadelphia, Pennsylvania.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E. Ho, and Joel Niklaus. 2023. SCALE: Scaling up the Complexity for Advanced Language Model Evaluation. *arXiv preprint*. ArXiv:2306.09237 [cs].

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, and *et al.* 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024. Qwen2. 5: A party of foundation models. *Qwen (Sept. 2024).*, 5.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Olivia Vaudaux, Caroline Bazzoli, Maximin Coavoux, Géraldine Vial, and Étienne Vergès. 2023. Pretrained language models v. court ruling predictions: A case study on a small dataset of French court of appeal rulings. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 38–43, Singapore. Association for Computational Linguistics.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Eva Wiesmann. 2019. Machine translation in the field of law: A study of the translation of Italian legal texts into German. *Comparative Legilinguistics*, 37(1):117–153.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications and opportunities. *Preprint*, arXiv:2408.07666.

Ran Zhang, Wei Zhao, and Steffen Eger. 2024. How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs. *arXiv preprint arXiv:2410.18697*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*. ArXiv: 1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9701–9708. AAAI Press.

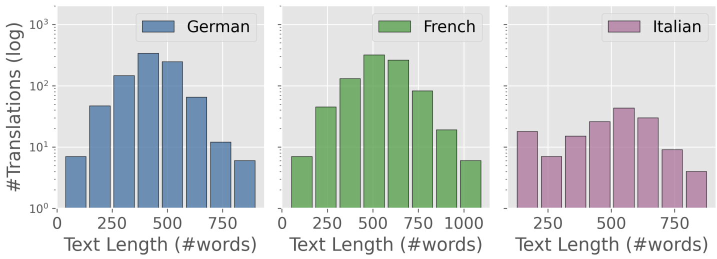
Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2765–2781. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus V1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

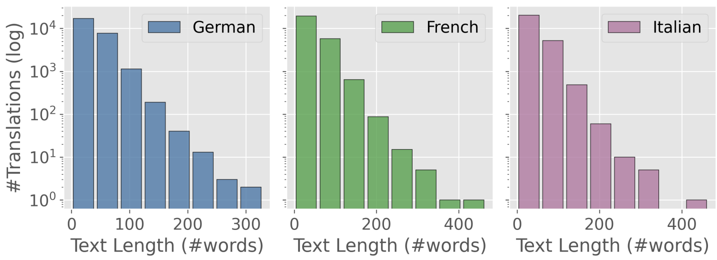
A Use of AI Assistants

We used GPT-4o and Claude Sonnet 3.5 for coding, shortening texts and editing LaTeX more efficiently.

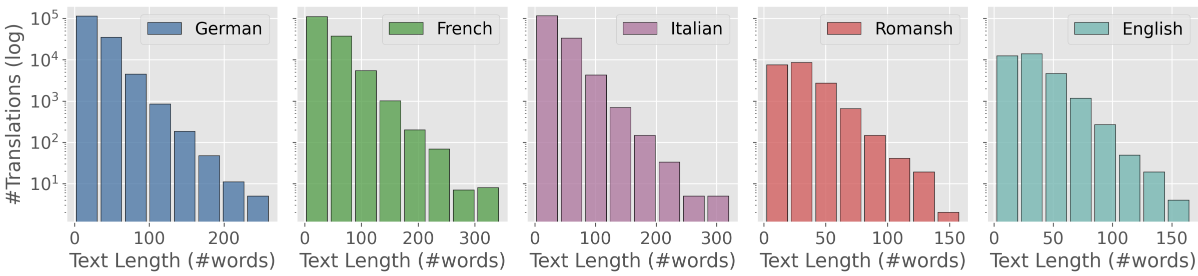
B Corpus Distribution of Text Lengths



(a) CH-Press-Trans dataset.

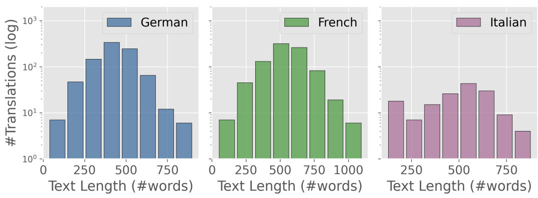


(b) CH-Headnote-Trans dataset (Text-Level).

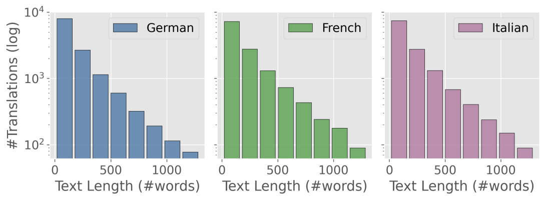


(c) CH-Law-Trans dataset (Paragraph-Level).

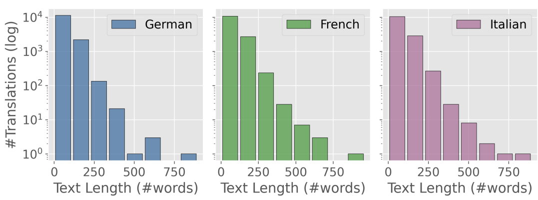
Figure B.1: SwiLTra-Bench text length distribution (training set).



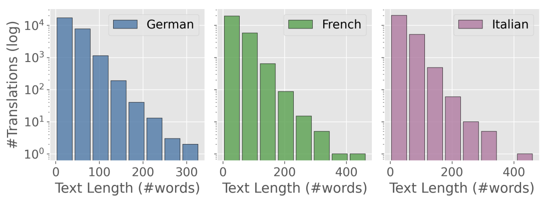
(a) CH-Press-Trans dataset.



(b) CH-Headnote-Trans dataset (BGE-Level).

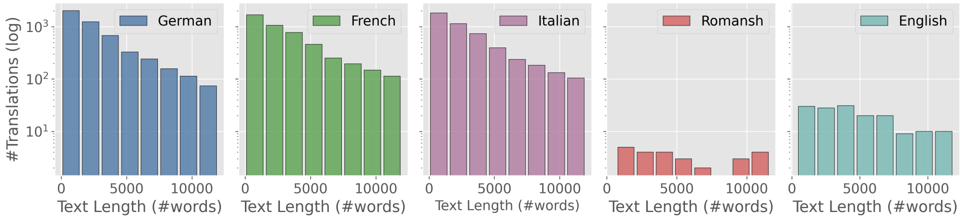


(c) CH-Headnote-Trans dataset (Regeste-Level).

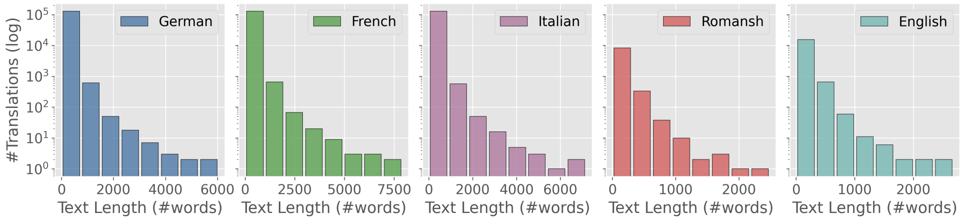


(d) CH-Headnote-Trans dataset (Text-Level).

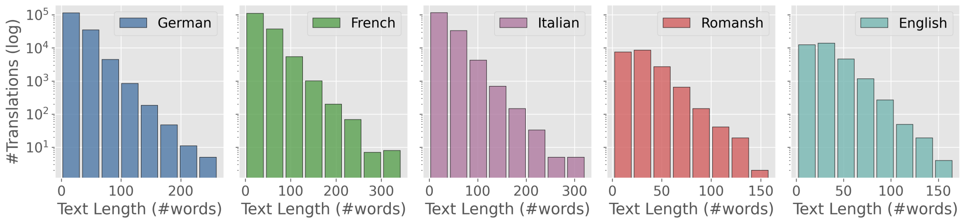
Figure B.2: Text length distribution of CH-Press-Trans and CH-Headnote-Trans dataset (training set).



(a) CH-Law-Trans dataset (Law-Level).



(b) CH-Law-Trans dataset (Article-Level).



(c) CH-Law-Trans dataset (Paragraph-Level).

Figure B.3: Text length distribution of CH-Law-Trans dataset (training set).

C Additional Experimental Results

Model	Family	Category	Size	↑ GEMBA-MQM	↑ XCOMET	↑ BLEURT	↑ METEOR	↑ ChrF
Gemma-2-2B	Gemma	open	2B	9.90 ± 0.1	35.52 ± 0.1	-102.08 ± 0.3	6.97 ± 0.1	11.12 ± 0.1
SLT-Gemma-2-2B	Gemma	finetuned	2B	72.71 ± 0.2	82.39 ± 0.1	26.72 ± 0.3	61.52 ± 0.1	79.48 ± 0.1
Gemma-2-9B	Gemma	open	9B	12.15 ± 0.1	36.54 ± 0.1	-102.19 ± 0.3	7.48 ± 0.1	0.00 ± 0.1
SLT-Gemma-2-9B	Gemma	finetuned	9B	82.54 ± 0.1	87.62 ± 0.1	32.89 ± 0.2	**65.16 ± 0.1**	78.95 ± 0.1
Llama-3.2-1B	Llama	open	1B	27.23 ± 0.2	48.43 ± 0.2	-15.64 ± 0.2	29.70 ± 0.1	39.87 ± 0.1
SLT-Llama-3.2-1B	Llama	finetuned	1B	64.14 ± 0.2	76.40 ± 0.1	26.35 ± 0.2	59.03 ± 0.1	79.76 ± 0.1
Llama-3.2-3B	Llama	open	3B	54.13 ± 0.2	67.43 ± 0.2	0.59 ± 0.2	38.57 ± 0.1	50.47 ± 0.1
SLT-Llama-3.2-3B	Llama	finetuned	3B	75.56 ± 0.2	83.39 ± 0.1	30.47 ± 0.2	62.54 ± 0.1	79.32 ± 0.1
Llama-3.1-8B	Llama	open	8B	67.09 ± 0.2	75.03 ± 0.2	6.25 ± 0.2	43.72 ± 0.1	50.56 ± 0.1
SLT-Llama-3.1-8B	Llama	finetuned	8B	80.23 ± 0.1	86.04 ± 0.1	31.91 ± 0.2	64.17 ± 0.1	**80.89 ± 0.1**
Phi-3.5-mini	Phi	open	3.8B	17.96 ± 0.2	41.93 ± 0.1	-92.40 ± 0.3	9.66 ± 0.1	11.42 ± 0.1
SLT-Phi-3.5-mini	Phi	finetuned	3.8B	73.90 ± 0.2	80.31 ± 0.1	10.33 ± 0.2	56.75 ± 0.1	76.72 ± 0.1
Phi-3-medium	Phi	open	14B	21.33 ± 0.2	38.91 ± 0.1	-81.04 ± 0.3	13.45 ± 0.1	17.74 ± 0.1
SLT-Phi-3-medium	Phi	finetuned	14B	81.56 ± 0.1	87.38 ± 0.1	32.39 ± 0.2	64.16 ± 0.1	80.40 ± 0.1
Qwen2.5-0.5B	Qwen	open	0.5B	9.82 ± 0.2	41.36 ± 0.2	-61.96 ± 0.2	14.53 ± 0.1	28.21 ± 0.1
SLT-Qwen2.5-0.5B	Qwen	finetuned	0.5B	52.37 ± 0.2	69.48 ± 0.2	22.67 ± 0.2	56.77 ± 0.1	78.66 ± 0.1
Qwen2.5-1.5B	Qwen	open	1.5B	35.21 ± 0.2	58.12 ± 0.2	-46.89 ± 0.3	22.55 ± 0.1	36.26 ± 0.1
SLT-Qwen2.5-1.5B	Qwen	finetuned	1.5B	69.58 ± 0.2	80.43 ± 0.1	26.90 ± 0.2	60.45 ± 0.1	78.13 ± 0.1
Qwen2.5-3B	Qwen	open	3B	48.85 ± 0.2	61.18 ± 0.2	-11.88 ± 0.2	34.77 ± 0.1	44.46 ± 0.1
SLT-Qwen2.5-3B	Qwen	finetuned	3B	74.33 ± 0.2	82.99 ± 0.1	27.78 ± 0.2	61.61 ± 0.1	78.03 ± 0.1
Qwen2.5-7B	Qwen	open	7B	58.79 ± 0.2	69.07 ± 0.2	0.73 ± 0.2	39.41 ± 0.1	42.67 ± 0.1
SLT-Qwen2.5-7B	Qwen	finetuned	7B	78.96 ± 0.1	86.03 ± 0.1	31.07 ± 0.2	63.40 ± 0.1	77.54 ± 0.1
Qwen2.5-14B	Qwen	open	14B	72.70 ± 0.2	79.54 ± 0.1	9.27 ± 0.2	45.06 ± 0.1	56.13 ± 0.1
SLT-Qwen2.5-14B	Qwen	finetuned	14B	**82.78 ± 0.1**	87.46 ± 0.1	32.37 ± 0.2	64.73 ± 0.1	78.56 ± 0.1
Qwen2.5-32B	Qwen	open	32B	70.30 ± 0.2	76.34 ± 0.1	6.91 ± 0.2	45.33 ± 0.1	57.94 ± 0.1
SLT-Qwen2.5-32B	Qwen	finetuned	32B	82.77 ± 0.1	**87.90 ± 0.1**	**33.20 ± 0.2**	65.07 ± 0.1	76.75 ± 0.1

Table C.1: Base models and their finetuned versions across different families and sizes.

D Corpus Examples

Dataset	Field	Comment
CH-Press-Trans	filename	Unique identifier of each press release
	de_text	Press release content in German
	fr_text	Press release content in French
	it_text	Press release content in Italian
	has_all_langs	Binary indicator of language availablity
CH-Law-Trans	abbreviation	The abbreviation of the law
	url	URL linking to the legal text on Fedlex
	rsNr	Swiss federal register number
	artNr	Article number
	parNr	Paragraph number
	dateApplicability	Date of applicability of the law
	{de/fr/it/rm/en}_lawTitle	Law titles in different languages
	{de/fr/it/rm/en}_artTitle	Article titles in different languages
	{de/fr/it/rm/en}_lawText	Full law texts in different languages
	{de/fr/it/rm/en}_artText	Full article texts in different languages
	{de/fr/it/rm/en}_parText	Full paragraph texts in different languages
	{de/fr/it/rm/en}_lawHtml	Law texts in HTML format in different languages
	{de/fr/it/rm/en}_artHtml	Article texts in HTML format in different languages
	{de/fr/it/rm/en}_parHtml	Paragraph texts in HTML format in different languages
CH-Headnote-Trans	bge	Case identifier
	year	Year of the court decision
	volume	Volume number of the court decision
	pageNumber	Page number of the court decision
	regesteNumber	Number assigned to the regeste
	textNumber	Number assigned to the specific text extract
	{de/fr/it}_bgeText	Full summary texts in different languages
	{de/fr/it}_regesteText	Regeste texts in different languages
	{de/fr/it}_regesteTitle	Regeste title in different languages
	{de/fr/it}_text	Text extract in different languages

Table D.1: Structure of the three SwiLTra-Bench datasets. Parallel translations for Romansh and English are only available in parts of the CH-Law-Trans dataset.

```
{
    'de_abbreviation': BV,
    'de_artText': Das Schweizervolk und die Kantone Zürich, Bern, Luzern, Uri,
        Schwyz, Obwalden und Nidwalden, Glarus, Zug, Freiburg, Solothurn,
        Basel-Stadt und Basel-Landschaft, Schaffhausen, Appenzell Ausserrhoden
        und Appenzell Innerrhoden, St. Gallen, Graubünden, Aargau, Thurgau,
        Tessin, Waadt, Wallis, Neuenburg, Genf und Jura bilden die
        Schweizerische Eidgenossenschaft.,
    ...
    'de_artTitle': Art. 1 Schweizerische Eidgenossenschaft,

    'fr_abbreviation': Cst.,
    'fr_artText': Le peuple suisse et les cantons de Zurich, de Berne, de
        Lucerne, d'Uri, de Schwyz, d'Obwald et de Nidwald, de Glaris, de Zoug,
        de Fribourg, de Soleure, de Bâle-Ville et de Bâle-Campagne, de
        Schaffhouse, d'Appenzell Rhodes-Extérieures et d'Appenzell
        Rhodes-Intérieures, de Saint-Gall, des Grisons, d'Argovie, de Thurgovie,
        du Tessin, de Vaud, du Valais, de Neuchâtel, de Genève et du Jura
        forment la Confédération suisse.
    ...
    'fr_artTitle': Art. 1 Confédération suisse,

    'it_abbreviation': Cost.,
    'it_artText': Il Popolo svizzero e i Cantoni di Zurigo, Berna, Lucerna, Uri,
        Svitto, Obvaldo e Nidvaldo, Glarona, Zugo, Friburgo, Soletta, Basilea
        Città e Basilea Campagna, Sciaffusa, Appenzello Esterno e Appenzello
        Interno, San Gallo, Grigioni, Argovia, Turgovia, Ticino, Vaud, Vallese,
        Neuchâtel, Ginevra e Giura costituiscono la Confederazione Svizzera.
    ...
    'it_artTitle': Art. 1 Confederazione Svizzera,


    'rm_abbreviation': Cst.,
    'rm_artText': Il pievel svizzer ed ils chantuns Turitg, Berna, Lucerna, Uri,
        Sviz, Sursilvania e Sutsilvania, Glaruna, Zug, Friburg, Soloturn,
        Basilea-Citad e Basilea-Champagna, Schaffusa, Appenzell Dadens ed
        Appenzell Dador, Son Gagl, Grischun, Argovia, Turgovia, Tessin, Vad,
        Vallais, Neuchâtel, Genevra e Giura furman la Confederaziun svizra.,
    ...
    'rm_artTitle': Art. 1 Confederaziun svizra,

    'en_abbreviation': Cst.,
    'en_artText': The People and the Cantons of Zurich, Bern, Lucerne, Uri,
        Schwyz, Obwalden and Nidwalden, Glarus, Zug, Fribourg, Solothurn, Basel
        Stadt and Basel Landschaft, Schaffhausen, Appenzell Ausserrhoden and
        Appenzell Innerrhoden, St. Gallen, Graubünden, Aargau, Thurgau, Ticino,
        Vaud, Valais, Neuchâtel, Geneva, and Jura form the Swiss Confederation.,
    ...
    'en_artTitle': Art. 1 The Swiss Confederation,
}
```

Listing 1: An Example of CH-Law-Trans: Article Dataset

```
{
    'bge': 100-IA-231,
    'year': 100,
    'volume': IA,
    'pageNumber': 231,

    'de_bgeText': Art. 85 lit. a OG. Ungültigerklärung einer kommunalen
        Volksinitiative wegen materieller Unvereinbarkeit mit dem kantonalen
        Recht. 1. Wieweit muss die Behörde beim Entscheid über die Gültigkeit
        einer kommunalen Initiative berücksichtigen, dass deren materielle
        Widerrechtlichkeit durch Annahme eines gleichzeitig eingereichten
        kantonalen Volksbegehrens dahinfallen könnte? (Erw. 2). 2. Die
        Verkehrsbetriebe der Stadt Zürich sind eine zur Eigenwirtschaftlichkeit
        verpflichtete ''produktive Unternehmun'' im Sinne von par. 129 des
        kantonalen Gemeindegesetzes. Die stadtzürcherische
        ''Gratistram-Initiative'', mit welcher ein grundsätzlicher Verzicht auf
        die Erhebung von Benützungsgebühren gefordert wurde, durfte daher wegen
        Unvereinbarkeit mit dem kantonalen Recht für ungültig erklärt werden
        (Erw. 3).,

    'fr_bgeText': Art. 85 lit. a OJ. Décision niant la validité d'une initiative
        communale en raison de son incompatibilité matérielle avec le droit
        cantonal. 1. Dans quelle mesure l'autorité qui se prononce sur la
        validité d'une initiative communale doit-elle tenir compte du fait que
        le contenu de cette dernière, contraire au droit, pourrait ne plus
        l'être en raison de l'acceptation d'une initiative cantonale déposée
        simultanément? (consid. 2). 2. Les entreprises de transport de la ville
        de Zurich, qui doivent être gérées selon les principes de l'économie
        industrielle, sont une ''entreprise à caractère productif'' au sens de
        l'art. 129 de la loi cantonale sur les communes. L'initiative communale
        zurichoise ''Gratistram'', qui exigeait en principe la suppression de
        toute taxe d'utilisation, pouvait être déclarée non valable en raison de
        son incompatibilité avec le droit cantonal (consid. 3).,

    'it_bgeText': Art. 85 lett. a OG. Diniego della validità di un'iniziativa
        comunale a causa della sua incompatibilità con il diritto cantonale. 1.
        In quale misura l'autorità che si pronuncia sulla validità di una
        iniziativa comunale deve tener conto del fatto che il contenuto di
        quest'ultima, contrario alla legge, cesserebbe d\'esserlo ove fosse
        accettata una iniziativa cantonale presentata nello stesso tempo?
        (consid. 2). 2. Le imprese di trasporto della città di Zurigo
        costituiscono una ''azienda produttiva'' ai sensi dell'art. 129 della
        legge cantonale sui comuni, tenuta come tale ad un esercizio secondo
        criteri economici. L'iniziativa comunale zurighese per il tram gratuito,
        che esigeva in linea di principio la soppressione d'ogni tassa
        d'utilizzazione, poteva quindi essere dichiarata invalida per la sua
        incompatibilità con il diritto cantonale (consid. 3).
}
```

Listing 2: An Example of CH-Headnote-Trans: BGE Dataset

```
{
    'de_text': ... Das BJ wies zuerst das Gesuch und dann die gegen diese
        Verfügung erhobene Einsprache des Betroffenen ab. Das
        Bundesverwaltungsgericht hiess die Beschwerde des Betroffenen gut, hob
        den Einspracheentscheid des BJ auf und wies die Angelegenheit dem BJ
        zurück, wogegen das BJ beim Bundesgericht eine Beschwerde eingereicht
        hat.

    Das Bundesgericht weist die Beschwerde ab. Gestützt auf eine vertiefte
        Auslegung des AFZFG kommt das Bundesgericht zum Schluss, dass ein Kind
        auch nach einer Adoption durch seine vormaligen Pflegeeltern als
        fremdplatziert im Sinne von Artikel 2 Buchstabe b des AFZFG gilt, womit
        es auch nach der Adoption von einer Fremdplatzierung betroffen ist und
        die Opfereigenschaft nach Artikel 2 Buchstabe d AFZFG erfüllen kann.

    'fr_text': ... L'OFJ a rejeté tant la demande que l'opposition formées par
        l'intéressé. Le Tribunal administratif fédéral a admis le recours de
        l'intéressé, annulé la décision sur opposition de l'OFJ et renvoyé
        l'affaire à l'OFJ, lequel a déposé un recours auprès du Tribunal fédéral.

    Le Tribunal fédéral rejette le recours. Sur la base d'une interprétation
        approfondie de la LMCFA, il parvient à la conclusion qu'un enfant doit ê
        tre considéré comme ayant fait l'objet d'un placement extrafamilial au
        sens de l'article 2 lettre b LMCFA même après avoir été adopté par ses
        parents nourriciers, si bien que la qualité de personne concernée et le
        statut de victime aux termes de l'article 2 lettre d LMCFA doivent lui ê
        tre reconnus même après l'adoption.

    'it_text': ... L'UFG ha respinto prima la domanda e poi l'opposizione
        interposta dall'interessato contro questa decisione. Il Tribunale
        amministrativo federale ha accolto il ricorso dell'interessato, ha
        annullato la decisione su opposizione resa dall'UFG e ha rinviato la
        questione all'UFG, che ha presentato ricorso al Tribunale federale.

    Il Tribunale federale respinge il ricorso. Sulla base di un'interpretazione
        approfondita della LMCCE, il Tribunale federale giunge alla conclusione
        che si deve ritenere che un bambino ha subito un collocamento
        extrafamiliare ai sensi dell'articolo 2 lettera b LMCCE anche dopo
        essere stato adottato dai genitori affilianti ed è pertanto riconosciuto
        come persona oggetto di misure nonché vittima secondo l'articolo 2
        lettera d LMCCE anche dopo l'adozione.
}
```

Listing 3: An Example of CH-Press-Trans Dataset

E Judge Correlations

Metric	Spearman (Bootstrap)	Spearman (CV)	RMSE (CV)	MAE (CV)
judge-ensemble	0.536 [0.453, 0.608]	0.533 ± 0.080	16.090 ± 1.424	12.979 ± 0.882
gemini-1-5-flash-codebook-diverse-deduction	0.506 [0.424, 0.586]	0.504 ± 0.074	15.215 ± 2.216	11.100 ± 0.670
XCOMET-XXL	0.484 [0.403, 0.567]	0.477 ± 0.093	14.877 ± 1.372	10.204 ± 0.748
gpt-4o-mini-codebook-single-deduction	0.474 [0.387, 0.551]	0.477 ± 0.095	22.168 ± 3.064	16.944 ± 1.691
gemini-1-5-flash-codebook-single-deduction	0.461 [0.374, 0.542]	0.466 ± 0.069	16.049 ± 2.212	10.990 ± 0.720
gpt-4o-mini-codebook-diverse-deduction	0.459 [0.378, 0.538]	0.459 ± 0.094	22.527 ± 2.678	17.138 ± 1.113
gpt-4o-codebook-single-deduction	0.444 [0.352, 0.538]	0.447 ± 0.070	29.020 ± 3.982	19.606 ± 1.951
gpt-4o-codebook-diverse-deduction	0.427 [0.334, 0.516]	0.412 ± 0.044	30.902 ± 1.843	21.537 ± 0.555
gpt-4o-detailed-single-absolute	0.418 [0.330, 0.501]	0.427 ± 0.052	35.940 ± 4.207	24.286 ± 3.382
gpt-4o-mini-basic-single-absolute	0.413 [0.320, 0.497]	0.422 ± 0.067	26.736 ± 1.817	21.370 ± 2.036
gpt-4o-basic-single-absolute	0.411 [0.320, 0.499]	0.411 ± 0.131	20.655 ± 2.947	14.596 ± 1.976
gpt-4o-detailed-diverse-absolute	0.378 [0.284, 0.464]	0.380 ± 0.090	34.786 ± 3.711	22.995 ± 3.173
gpt-4o-mini-detailed-single-absolute	0.378 [0.282, 0.471]	0.384 ± 0.069	35.437 ± 1.866	30.302 ± 1.718
gpt-4o-basic-diverse-absolute	0.376 [0.281, 0.467]	0.383 ± 0.087	21.780 ± 2.249	13.550 ± 1.382
gpt-4o-mini-detailed-diverse-absolute	0.358 [0.259, 0.450]	0.364 ± 0.097	36.480 ± 2.365	30.828 ± 2.387
bleurt_large	0.356 [0.261, 0.445]	0.364 ± 0.147	63.110 ± 5.225	58.102 ± 5.064
gpt-4o-mini-basic-diverse-absolute	0.354 [0.263, 0.447]	0.361 ± 0.048	28.363 ± 1.315	22.393 ± 1.347
gemini-1-5-pro-basic-single-absolute	0.306 [0.209, 0.404]	0.295 ± 0.083	36.203 ± 3.618	22.993 ± 3.010
gemini-1-5-pro-codebook-diverse-deduction	0.303 [0.200, 0.397]	0.298 ± 0.095	34.528 ± 3.882	20.580 ± 2.477
GEMBA-MQM_gpt-4o	0.293 [0.189, 0.383]	0.289 ± 0.093	18.331 ± 1.743	12.698 ± 0.787
gemini-1-5-pro-codebook-single-deduction	0.292 [0.195, 0.392]	0.292 ± 0.074	36.718 ± 2.663	21.816 ± 2.158
gemini-1-5-flash-detailed-single-absolute	0.281 [0.190, 0.375]	0.275 ± 0.049	29.067 ± 3.070	18.709 ± 1.651
gemini-1-5-flash-basic-single-absolute	0.270 [0.179, 0.359]	0.279 ± 0.110	33.283 ± 5.239	20.412 ± 3.287
gemini-1-5-flash-basic-diverse-absolute	0.255 [0.157, 0.351]	0.249 ± 0.069	27.649 ± 5.852	16.756 ± 3.233
gemini-1-5-pro-basic-diverse-absolute	0.252 [0.153, 0.351]	0.250 ± 0.082	36.544 ± 3.470	22.990 ± 2.684
gemini-1-5-pro-detailed-diverse-absolute	0.247 [0.147, 0.345]	0.250 ± 0.097	38.316 ± 2.063	26.098 ± 1.148
gemini-1-5-pro-detailed-single-absolute	0.235 [0.137, 0.334]	0.244 ± 0.079	38.618 ± 2.362	27.798 ± 1.977
gemini-1-5-flash-detailed-diverse-absolute	0.231 [0.130, 0.328]	0.225 ± 0.055	30.091 ± 4.891	19.671 ± 3.346
BERTScore-F	0.163 [0.066, 0.268]	0.170 ± 0.053	36.723 ± 1.340	31.523 ± 1.772
meteor	0.160 [0.057, 0.259]	0.164 ± 0.125	34.170 ± 3.444	29.270 ± 3.191

Table E.1: Correlation metrics with human scores (with 95% CIs and Cross-Validation)

F Judge Prompts

System Prompt

```
Act as a Judge specializing in the evaluation of translations of Swiss legal
    documents. Your task is to assess the accuracy, clarity, and fidelity of the
    model's translation to the golden translation, while considering the nuances
    of legal language.
```

User Prompt

```
You will be provided with a source text, its golden translation, and the model's
    translation. Your task is to judge how correct the model's translation is
    based on the golden translation, and then give a correctness score. The
    correctness score should be one of the below numbers: 0.0 (totally wrong),
    0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 (totally right). You
    should first briefly give your reasoning process regarding how the model's
    translation conforms to or contradicts the golden translation, and then give
    the correctness score. The correctness score must strictly follow this
    format: \"[[score]]\", e.g., \"The correctness score: [[0.5]]\". Below are
    some examples.
```

Listing 4: The system and user prompt of the *basic* judge setup.

G Annotation Guidelines

System Prompt

```
You are a senior legal translator and quality assurance specialist with over 20 years of experience in Swiss law, certified by the Swiss Sworn Translators Association (Association suisse des traducteurs-jurés, ASTJ). You possess native-level proficiency in all Swiss national languages (German, French, Italian, and Romansh) as well as English, enabling precise evaluation of legal nuances across all linguistic combinations. Your task is to evaluate machine-translated legal texts for accuracy, clarity and fidelity to Swiss legal standards analyzing the subtle complexities of legal language. You excel at identifying even minor discrepancies and calibrating evaluation scores appropriately to reflect the severity of each error.
```

User Prompt

```
INPUT FORMAT:
Source Text: [Original text in source language]
Golden Translation: [Reference professional translation]
Model Translation: [Machine-generated translation to be evaluated]

EVALUATION DIMENSIONS:
Accuracy: Semantic equivalence, correct legal terminology, and preservation of legal meaning.
Clarity: Logical flow, appropriate legal register, and unambiguous expression.
Fidelity: Adherence to Swiss legal conventions, jurisdiction-specific terminology, and formal register.

SCORING RUBRIC:
1.0: Perfect translation
0.7-0.9: Minor issues only
0.4-0.6: Significant but non-critical errors
0.1-0.3: Major errors affecting legal meaning
0.0: Completely incorrect

EVALUATION GUIDELINES:
Stylistic differences should not impact accuracy significantly unless they alter the legal meaning.
Untranslated Latin terms (e.g., prima facie) are not considered errors, but they should still be assessed for appropriate use within the context of the answer.
Terminology should be used consistently throughout the text.
Consider both explicit and implicit legal meanings.
Consider jurisdiction-specific legal terminology.
Flag any ambiguities, omissions or additions that affect legal meaning.

REQUIRED OUTPUT FORMAT:
Your response should be in plain text with the following sections:
Reasoning: Analyze how the model's translation aligns with or differs from the golden translation, focusing on significant legal and linguistic aspects.
Examples: Identify specific terms, phrases, or sections in the model's answer that were correct or incorrect, with explanations.
Score: End with exactly this format: \"The correctness score: [[score]]\"
The correctness score must strictly follow this format: \"[[score]]\", e.g., \"The correctness score: [[0.5]]\". Below are some examples.
```

Listing 5: The system and user prompt of the *detailed* judge setup.

System Prompt

```
You are a senior legal translator and quality assurance specialist with over 20 years of experience in Swiss law, certified by the Swiss Sworn Translators Association (Association suisse des traducteurs-jurés, ASTJ). You possess native-level proficiency in all Swiss national languages (German, French, Italian, and Romansh) as well as English, enabling precise evaluation of legal nuances across all linguistic combinations. Your task is to evaluate machine-translated legal texts for accuracy, clarity and fidelity to Swiss legal standards analyzing the subtle complexities of legal language. You excel at identifying even minor discrepancies and calibrating evaluation scores appropriately to reflect the severity of each error.
```

User Prompt

```
GENERAL INSTRUCTIONS:
You must give each translation a score between 0 and 1 that must be divisible by 0.1 (e.g., 0.6 or 0.9). To this end, you are given a source text, its ''gold translation'' (official translation of the Swiss authorities) and the predicted translation, to which you must assign the score. You can also write down notes if deemed necessary.

SCORE:
The scores shall reflect the completeness and accuracy of the predicted translation. In other words, you should not give a score based on readability or stylistic attributes.

POINT DEDUCTION SYSTEM:
A perfect, i.e., a perfectly complete and accurate translation receives a score of 1.
0.1 points deduction for a relevant legal term in an unusual but still correct manner. 0.1 points shall also be deducted if the law has not been translated (e.g., BV to BV). Finally, 0.1 points shall be deducted if a non-relevant term is missing.
0.2 points deduction if a legally relevant legal term is translated erroneously. 0.2 points shall also be deducted if a relevant term is missing.
0.4 points deduction for critical errors, such as when a law is translated with reference to the wrong law.

Do not deduct points for discrepancies between the predicted translation and the gold translation if the predicted translation matches the source text better. The gold translation should primarily serve as a reference to help you assess cases where it is also a correct translation of the source. In some cases, the source text may differ slightly from the gold translation. This can happen if the source text itself was previously translated. Repeated errors for the same term should not lead to multiple point deductions.

REQUIRED OUTPUT FORMAT:
Your response should be in plain text with the following sections:
Deductions: Focusing on significant legal and linguistic aspects, analyze and present concretely all points to be deducted together with brief explanations.
Score: End with exactly this format: \"The correctness score: [[score]]\"
The correctness score must strictly follow this format: \"[[score]]\", e.g., \"The correctness score: [[0.5]]\". Below are some examples.
```

Listing 6: The system and user prompt of the *codebook* judge setup.

General Instructions: Annotators must give each translation a score between 0 and 10 that must be divisible by 1 (e.g., 6 or 9). To this end, annotators are given a source text, its ''gold translation'' (official translation of the Swiss authorities) and the predicted translation, to which they must assign the score. Annotators can also write down notes if deemed necessary.

Score: The scores shall reflect the completeness and accuracy of the predicted translation. In other words, annotators should not give a score based on readability or stylistic attributes.

Point Deduction System: The scoring should be conducted using a points deduction scheme.

A perfect, i.e., a perfectly complete and accurate translation receives a score of 10.
1 points deduction for a relevant legal term in an unusual but still correct manner. 1 point shall also be deduced if the law has not been translated (e.g., BV to BV). Finally, 1 point shall be deduced if a non-relevant term is missing.
2 points deduction if a legally relevant legal term is translated erroneously. 2 points shall also be deduced if a relevant term is missing.
4 points deduction for critical errors, such as when a law is translated with reference to the wrong law. If a new category of critical error is introduced under this deduction, the annotator must inform the other annotators through their communication channel.

Do not deduct points for discrepancies between the predicted translation and the gold translation if the predicted translation matches the source text better. The gold translation should primarily serve as a reference to help you assess cases where it is also a correct translation of the source. In some cases, the source text may differ slightly from the gold translation. This can happen if the source text itself was previously translated.

Notes for Multiple Deductions: If two or more deductions are applied, annotators must briefly document the individual deductions in the comments field, e.g., ''-1, -1, -2''. This allows for potential adjustments to weighting later to account for text length if necessary. Repeated errors for the same term should not lead to multiple point deductions.

Subjectivity: We are aware that the scoring system is subject to a certain degree of subjectivity. However, assessing the quality of a translation cannot be fully objectified. To demonstrate how the scoring system works in practice, we provide annotators with 3 examples including a suggested score.

Examples:

1) Source: ''Bewilligungen nach diesem Artikel dürfen nur erteilt werden, wenn:''
Gold: ''Permits under this Article may be issued only if:''
Prediction: Permits under this Article may only be granted if:
Score: 10

2) Source: Bank client confidentiality and other client and professional confidentiality protected by law shall be maintained.
Gold: Das Bankgeheimnis und andere gesetzlich geschützte Kunden- und Berufsgeheimnisse sind zu wahren.
Prediction: Die gesetzlich geschützte Vertraulichkeit von Bankkundeninformationen sowie andere gesetzlich geschützte Kunden- und Berufsgeheimnisse sind zu wahren.
Score: 9 (-1 for unusual translation of ''Bankgeheimnis'')

3) Source: 1. La constitution de sûretés par la partie adverse (art. 79 al. 2 LBI) ne dispense pas le juge d'examiner s'il y a lieu d'ordonner des mesures provisionnelles aux conditions prévues à l'art. 77 al. 2 LBI.
Gold: 1. Eine Sicherheitsleistung gemäss Art. 79 Abs. 2 PatG enthebt den Richter nicht von der Prüfung der Frage, ob die Voraussetzungen für vorsorgliche Massnahmen nach Art. 77 Abs. 2 PatG gegeben seien.
Prediction: Die Stellung von Sicherheiten durch die Gegenpartei (Art. 79 Abs. 2 BEHG) entbindet den Richter nicht von der Prüfung, ob vorsorgliche Massnahmen unter den in Art. 77 Abs. 2 BEHG vorgesehenen Bedingungen anzuordnen sind.
Score: 6 (-4 for highly relevant erroneous translation of ''LBI'' to ''BEHG'' instead of ''PatG'')

Listing 7: The annotation guidelines given to the human experts.