LawInstruct: A Resource for Studying Language Model Adaptation to the Legal Domain

Anonymous ACL submission

Abstract

Instruction tuning is an important step in making language models useful for direct user interaction. However, the legal domain is underrepresented in typical instruction datasets (e.g., only 10 out of 1600+ tasks in Super-NaturalInstructions). To study whether instruction tuning on legal datasets is necessary for strong legal reasoning, we aggregate 58 annotated legal datasets and write instructions for each, creating LawInstruct. LawInstruct covers 17 global jurisdictions, 24 languages and a total of 12M examples across diverse tasks such as legal QA, summarization of court cases, and legal argument mining. We evaluate our models on LegalBench, measuring legal reasoning across five categories in 162 challenging and realistic legal tasks, and MMLU, to measure potential drops in general reasoning capabilities. We find that legal-specific instruction tuning on Flan-T5 – yielding FLawN-T5 – improves performance on LegalBench across all model sizes, with an aggregate increase of 15 points or 50% over Flan-T5 for the base size. No model size shows performance drops in MMLU. We publish LawInstruct as a resource for further study of instruction tuning in the legal domain.

1 Introduction

In recent years, Large Language Models (LLMs) advanced significantly, evident in their performance gains across numerous benchmarks, including SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021a), and various human examinations (OpenAI, 2023), such as the U.S. bar exams for law practice admission (Katz et al., 2023). However, the interplay between domain-specific training and within-domain evaluation is poorly understood. This work examines how training on domain-specific legal corpora affects performance on the widest set of legal-domain evaluation benchmarks known to the authors. We thus conduct a study of the ability of models to answer questions, classify, make judgments, extract information, and otherwise perform decision making or higher-order cognitive tasks (i.e., to "reason") within a limited domain, as opposed to broad-domain benchmarking. We present evidence that domain-specific pretraining and instruction tuning improve performance—but the effect does not generalize across all tasks, training regimes, model sizes, and other factors.

Although large closed models also still hallucinate heavily on legal texts (Dahl et al., 2024), they achieve much better performance on LegalBench than smaller open models (e.g., 77.3 for GPT-4 vs. 60.1 for Flan-T5 XXL, the state-of-the-art open model). In the legal domain it is often crucial for reasons of trust and data protection not to use public models, so many firms need on-premise deployments. Therefore models like Claude or GPT-4 cannot be used, stressing the need for open models. In this study, we explore the potential of enhancing model performance through in-domain instruction tuning and continued pretraining on Flan-T5, the current state-of-the-art open model on LegalBench in both the 3B and 11B range.

To study this, we use the MultiLegalPile (Niklaus et al., 2023b), a 689GB multilingual legal corpus, for continued pretraining. Because no instruction dataset for legal reasoning is available, we introduce LawInstruct, spanning 24 languages in 17 jurisdictions on four continents. It contains 12M training examples for QA, entailment, summarization, and information extraction tasks in the legal domain, each presented as a bespoke instruction with corresponding output. With this large instruction dataset in hand, we fine-tune models and then perform quantitative analyses of their outputs on the LegalBench (Guha et al., 2023) and MMLU (Hendrycks et al., 2021b) benchmark suites. Instruction tuning Flan-T5 models on LawInstruct, we achieve a balanced accuracy of 58.1 on LegalBench for the XL size, improving by 8 points or

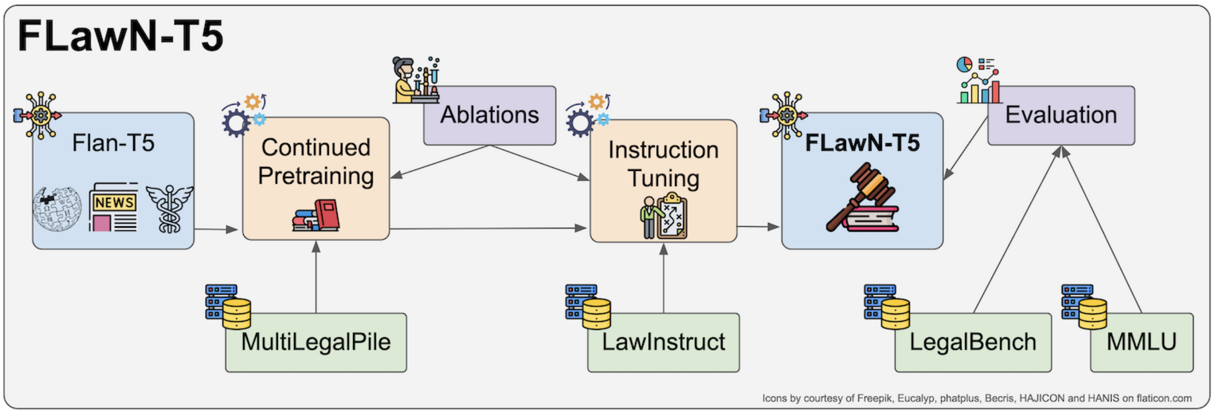


Figure 1: We continue pretraining on MultiLegalPile, instruction tune on LawInstruct and evaluate on LegalBench and MMLU.

16% over the baseline. The Small model even improves by 9.6 points or 38.1% and by 14 points or 55.4% when we also continue pretraining it.

The contributions of this paper are four-fold: First, we curate the first legal instruction dataset by standardizing and writing instructions for 58 high-quality annotated datasets covering diverse legal tasks to make them usable for instruction tuning in the first place. Second, we continue pretraining and instruction tune T5, mT5, and Flan-T5 models and achieve new state-of-the-art on LegalBench in all tested parameter ranges. Third, we perform a wide range of ablations across different dataset configurations deepening our understanding of adapting models to specific domains. Finally, we publicly release the permissively-licensed portion of the curated dataset on the Hugging Face Hub[1] and release the code used to create the dataset[2] including pointers on how to access the portions of the data that require special agreements.

2 Experimental Setup

In this section, we describe the experimental setup we used to test the effect of pretraining and instruction tuning on in-domain legal data. We use random seed 42 throughout. Our experiments were performed with T5X [3] on TPUv4 pods using 2 to 512 cores. We present the mean across tasks per LegalBench category and for LegalBench overall by aggregating over the categories. We consider T5 v1.1+LM adaptation (Raffel et al., 2020; Lester et al., 2021), Flan-T5 (Chung et al., 2022) and mT5 (Xue et al., 2021) models in the sizes Small, Base, XL and XXL, allowing us to study effects over different model scales. We selected the T5 family of models over other models for three reasons: 1) Flan-T5 XL and XXL perform best in their parameter range on LegalBench, 2) T5 and mT5 allow us to measure the effect of multilinguality in a controlled setting, and 3) the T5 model family contains models from 60M parameters (Small) to 11B (XXL) allowing us to study scaling behaviour also at smaller scales.

[1] URL available upon acceptance

[2] URL available upon acceptance

[3] https://github.com/google-research/t5x

2.1 Continued Pretraining

We continue pretraining on the **MultiLegalPile** (Niklaus et al., 2023b), a 689GB corpus in 24 languages from 17 jurisdictions. It includes diverse legal data sources with varying licenses and allows for pretraining NLP models under fair use, with more permissive licenses for the Eurlex Resources and Legal mC4 subsets. It consists of four large subsets: a) Native Multi Legal Pile (112 GB), b) Eurlex Resources (179 GB), c) Legal mC4 (106 GB), and d) Pile of Law (292 GB). For our mT5 experiments, we use the entire corpus, and for T5 and Flan-T5 experiments, we use only English texts.

We continued pretraining (a.k.a. domain adaptation of) with 512 tokens in both inputs and targets on the MultiLegalPile (Niklaus et al., 2023b) whereas the original models were pretrained on C4 (Raffel et al., 2020). We used the UL2 mixture (Tay et al., 2022) due to its promise to enable improved training efficiency with its mixture of denoisers. In initial experiments we used batch size 1024 and warmed up the learning rate linearly for the first 10K steps from 2.5e-3 to 5e-3, then decayed it to 1.5e-3. However, we noticed training instabilities for the XXL models. We switched to a constant learning rate of 1e-3 and ran a sweep over batch sizes 64, 128, 256, 512, 1024. The XXL model trained stably only with batch size 128.

2.2 Instruction Tuning

In this paper, we are interested in the ability of LLMs to answer questions, make judgments, and perform decision making (i.e., to "reason") within the legal domain. Legal reasoning is often highly sensitive, and the struggles of factuality in LLMs lead to legalese with "bogus judicial decisions, bogus quotes, and bogus internal citations" (Weiser, 2023; Dahl et al., 2024). In the absence of legal instruction datasets and to evaluate the effect of legal instruction tuning on models' capability to reason in legal domains, we develop **LawInstruct**: a large instruction dataset that normalizes and adapts 58 existing or novel legal-domain datasets with custom templates. LawInstruct is the first instruction dataset in the legal domain known to the authors. We attempted to collect a broad sample of datasets to expose the model to a variety of legal systems and concepts. We started by taking the datasets operating on legal data from Natural Instructions (Mishra et al., 2022; Wang et al., 2022) and then surveyed the literature to select high-quality legal datasets. The resulting dataset contains a total of almost 12M examples in 24 languages. Data sources and detailed statistics including license, language and jurisdiction are given in Appendix B Table 2. Each example is built from a human-written task-specific template: We write a simple instruction per task (107 in total), we take the input of the supervised dataset as the prompt and the output as the answer (see Figure 2 for an example). We show pie charts visualizing the composition of LawInstruct across the jurisdiction and task types calculated by the number of examples in Figure 3 and by datasets in Appendix B Figure 11. Figure 4 shows the length distribution of the instructions, prompts and answers following Wang et al. (2023b).

We finetuned the models with 2048 input and 512 target tokens. We ran a hyperparameter sweep for the XL model over the learning rate (5e-5, 1e-5, 5e-4, 1e-4, 5e-4) and dropout (0, 0.05, 0.1, 0.15, 0.2, 0.25), with learning rate 5e-4 and dropout 0.15 achieving the best validation loss. Unless specified otherwise, we trained the models for 2K steps with batch size 64. In addition to LawInstruct, we used an updated Flan mixture (Chung et al., 2022). We built the input by concatenating the prompt with two new lines, the instruction and two additional new lines. Per LawInstruct config, we used the first 16 examples for validation and the remaining ones for training. We selected the model with the best LawInstruct validation loss. While in-context learning has achieved strong results in many tasks (Brown et al., 2020), further finetuning language models for specific tasks may still be necessary for better results (Mosbach et al., 2023).

Instruction: Answer the following legal question. Cite relevant evidence when possible.

Prompt: Question: So say a bank robber walks into a bank and hands the teller a note saying simply "please give me $1,000" (or maybe even just verbally saying "I need $1,000 please") and if the teller accepts, which they probably would because they're supposed to comply with the robber, he takes the money, otherwise if they refuse, he leaves. How is this different than say people asking for charity donations in the street? He just politely requested money without making any threats.

Answer: Answer: As a concrete example, consider Missouri v. Coleman, where Coleman handed a teller a plastic bag and said "I need you to do me a favor. Put the money in this bag", and later as the manager approached, said "Ma'am, stop where you are and don't move any farther". Coleman was convicted of second degree robbery: but the appeals court found that he had not acted forcibly, so his conviction was overturned. Instead, the court entered a conviction for the lesser offense of stealing which is when one appropriates property or services of another with the purpose to deprive him or her thereof, either without his or her consent or by means of deceit or coercion. The act may also be termed "theft", as in Washington state. The thief is acting deceptively and thereby gaining control over property.

Figure 2: Instruction template in LawInstruct for StackExchangeQuestionsLegal populated with *instruction*, *prompt* and *answer*. Models are trained to generate the answer conditioned on the instruction and prompt.

2.3 Evaluation

We evaluate our models on LegalBench and MMLU to test in-domain and generalization performance, respectively. **LegalBench** (Guha et al., 2023) consists of 162 tasks evaluating different aspects of legal classification and reasoning. Each task is assigned to one of five categories, depending on the broader type of legal reasoning implicated. LegalBench tasks are sourced from both previously constructed datasets and novel tasks collected from different members of the legal community (e.g., lawyers, legal impact organizations, legal academics). As such, LegalBench is thought to capture tasks of interest and practical applicability. LegalBench tasks span a wide range of legal subject areas (e.g., contracts, civil procedure, tax, etc.) and text-types (natural language, contractual terms, judicial opinions, etc.). The majority of tasks are either classification or extraction tasks, thus enabling automated evaluation. Massively

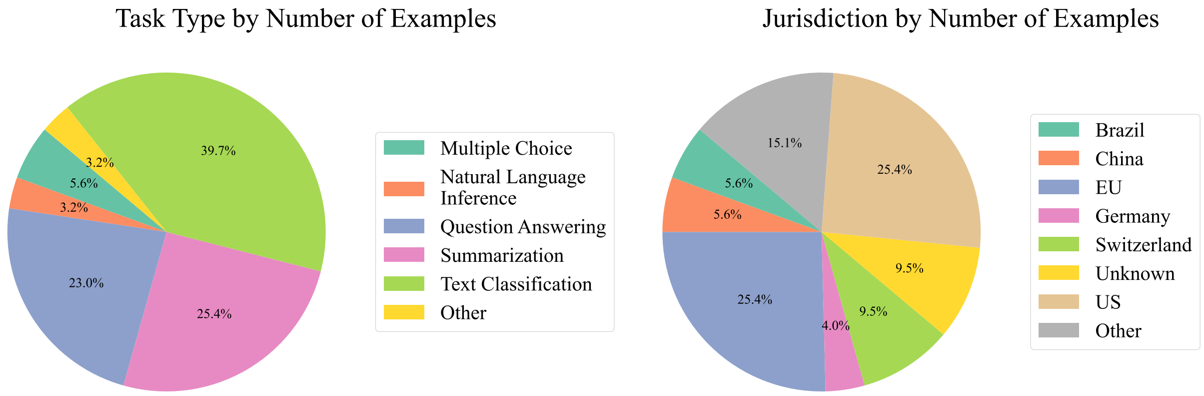
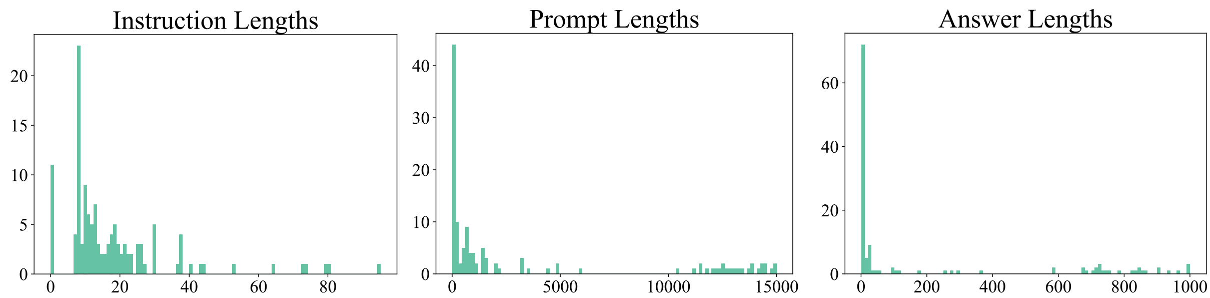


Figure 3: Jurisdiction and task type by examples.



(a) Instructions (capped at 100) (b) Prompts (capped at 15K) (c) Answers (capped at 1K)

Figure 4: Mean length distributions for instructions, prompts and answers.

Multilingual Language Understanding (**MMLU**) benchmarks models factual knowledge (Hendrycks et al., 2021b). MMLU contains multiple-choice questions on 57 subjects, including three related to law: jurisprudence, international law, and professional law. While multilingual benchmarks like LEXTREME (Niklaus et al., 2023a) exist, they remain challenging for generative models not fine-tuned per task. Therefore, we focus on LegalBench and MMLU, both in English.

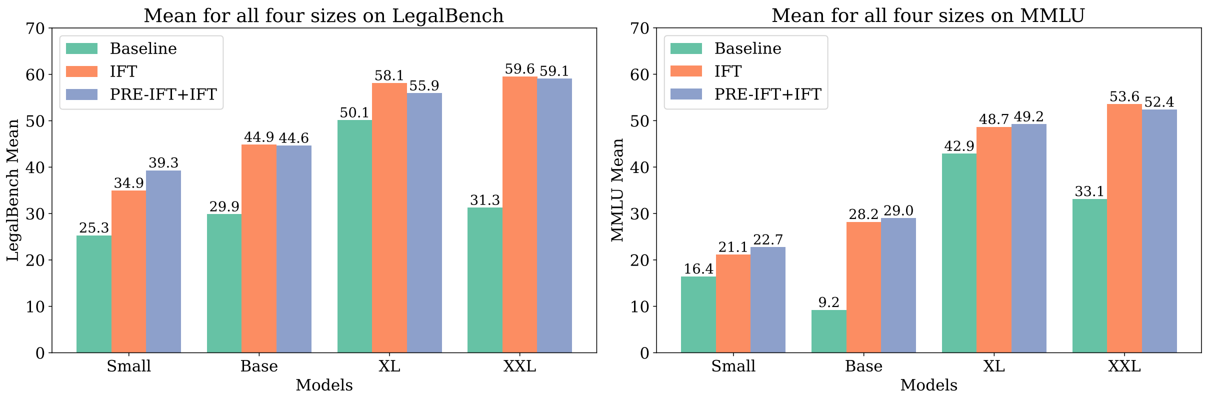
For evaluation, we set temperature to 0 in line with accepted practice for LegalBench evaluation (Guha et al., 2023) that focuses on the highest-likelihood token sequence with minimal variance. We removed the following prefixes before scoring: "label", "target", "option", "answer", "a:". We did not evaluate Rule QA because it necessitated manual evaluation. We show paper baseline results compared with our runs in Appendix E Table 5. Our XL model is quite close to the XL model in the LegalBench paper, but there are significant differences for the XXL model. We provide a more detailed analysis of possible causes in Appendix C.1. Unless specifically mentioned, we compare to our baselines results. We hold out LegalBench tasks overlapping with LawInstruct tasks unless specified otherwise (see Appendix C.2 for details).

3 Results

This section discusses the main results from instruction tuning and continued pretraining Flan-T5.

Figure 5 and Table 1 show the performance progression from the baseline over instruction tuning to domain adaptation + instruction tuning on LegalBench and MMLU. Instruction tuning leads to a large performance increase for all model sizes (38.1% for Small, 50.2% for Base, 16% for XL, and 90.5% for XXL). Domain adaptation + instruction tuning only improves further for the Small model size (55.4% vs. 38.1%). It seems like larger models benefit less from in-domain pretraining than smaller models, possibly because they can "remember" more from the pretraining phase due to increased capacity. Alternatively, a reason for non-consistent improvements of domain adaptation could be the switch from the UL2 tasks in continued pretraining to standard next-token prediction in instruction tuning. Finally, we conjecture that the switch from input length 512 tokens in continued pretraining to 2048 tokens in instruction tuning could have led lower performance for domain-adapted models.

To analyze the change in performance in more detail, we show the difference to the baseline for the XL model on LegalBench and MMLU across tasks (see Figure 6) and across categories (see Fig-



(a) LegalBench

(b) MMLU

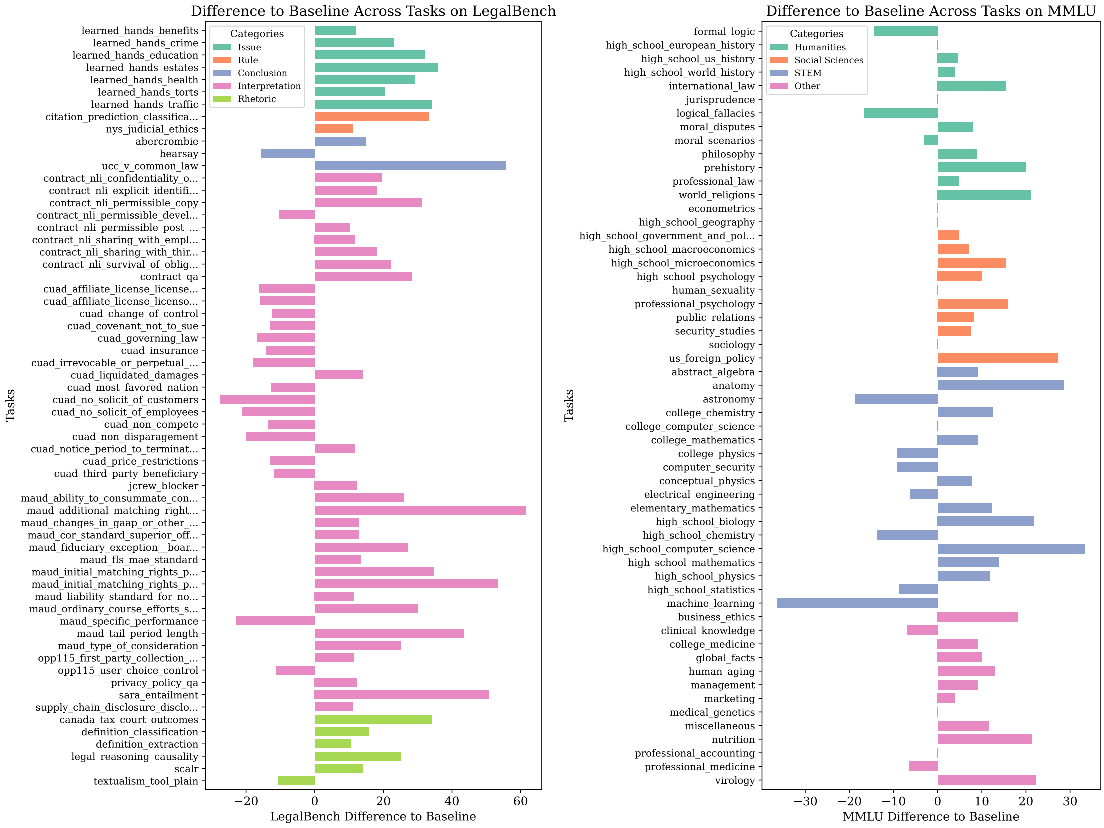
Figure 5: Performance progression on LegalBench and MMLU from baseline to instruction tuning (IFT) and continued pretraining followed by instruction tuning (PRE-IFT+IFT).

Table 1: Progression of performance from baseline to instruction tuning (IFT) and continued pretraining followed by instruction tuning (PRE-IFT+IFT).

LLM	**Issue**	**Rule**	**Conclusion**	**Interpretation**	**Rhetorical**	**LegalBench**	**Improvement**
Small Baseline	0.3 ± 0.7	30.4 ± 20.3	39.8 ± 20.8	28.2 ± 21.6	27.7 ± 21.9	25.3 ± 14.8	-
Small IFT	25.0 ± 22.0	38.1 ± 25.4	43.0 ± 17.1	36.1 ± 26.5	32.6 ± 24.2	34.9 ± 6.7	9.6 (38.1%)
Small PRE-IFT+IFT	51.6 ± 2.7	37.7 ± 25.2	39.8 ± 18.4	33.7 ± 23.3	33.8 ± 22.4	39.3 ± 7.4	14.0 (55.4%)
Base Baseline	44.7 ± 12.4	18.0 ± 23.6	20.9 ± 24.8	28.9 ± 21.2	37.0 ± 21.3	29.9 ± 11.1	-
Base IFT	50.3 ± 2.4	38.8 ± 25.9	40.5 ± 15.7	49.5 ± 19.1	45.2 ± 22.0	44.9 ± 5.2	15.0 (50.2%)
Base PRE-IFT+IFT	51.6 ± 4.8	38.2 ± 25.5	44.0 ± 13.4	45.4 ± 16.5	44.1 ± 19.0	44.6 ± 4.8	14.8 (49.5%)
XL Baseline	53.5 ± 6.0	32.1 ± 24.6	46.8 ± 15.6	58.7 ± 21.3	59.6 ± 25.6	50.1 ± 11.3	-
XL IFT	65.7 ± 15.2	45.1 ± 30.3	49.5 ± 14.2	61.7 ± 17.1	68.6 ± 24.1	58.1 ± 10.3	8.0 (16.0%)
XL PRE-IFT+IFT	60.3 ± 10.6	44.3 ± 29.7	50.5 ± 15.4	57.3 ± 15.9	67.3 ± 23.1	55.9 ± 8.9	5.8 (11.6%)
XXL Baseline	36.1 ± 21.5	18.8 ± 24.6	25.2 ± 26.0	35.1 ± 22.2	41.1 ± 18.4	31.3 ± 9.1	-
XXL IFT	55.2 ± 23.7	46.3 ± 31.6	56.2 ± 18.3	66.3 ± 19.7	73.8 ± 24.4	59.6 ± 10.6	28.3 (90.5%)
XXL PRE-IFT+IFT	52.2 ± 14.7	47.4 ± 30.8	59.2 ± 18.3	66.6 ± 18.5	70.0 ± 24.1	59.1 ± 9.5	27.8 (89.0%)
GPT-4 Guha et al. (2023)	82.9	59.2	89.9	75.2	79.4	77.3	-

ure 7). We find that FLawN-T5 outperforms baseline Flan-T5 in most LegalBench tasks in most categories. The exception are tasks in the interpretation category, specifically CUAD (Hendrycks et al., 2021c), where the fine-tuned model is actually worse than the baseline by around 10 points on average. A possible explanation could be negative transfer from the instruction tuning data since the task formulations are very different to the instructions in LegalBench. In MAUD (Wang et al., 2023a) and Contract-NLI (Koreeda and Manning, 2021), the instructions are much more similar from LawInstruct to LegalBench, leading to improvements compared to the baseline. On MMLU, most categories and tasks see increases in performance, especially the categories social sciences and other. We find that performance suffers mostly in the STEM category and to some extent in the humanities. Interestingly, the largest drop is in machine learning but the largest rise is in high school computer science. In the humanities, more "hard" disciplines are affected by performance decrease, such as formal logic and logical fallacies.

Across categories overall we see lower improvements in conclusion and interpretation. Conclusion is one of LegalBench categories requiring more sophisticated reasoning capabilities; maybe larger models would see larger gains there. Concurrent work (Colombo et al., 2024) instruction tuned on synthetic legal data. They even saw a drop in performance in conclusion tasks compared to the baseline arguing, that conclusion tasks "require much more pure deductive reasoning than actual legal knowledge" compared to tasks from the other categories. Lower improvement in interpretation could be explained by negative transfer caused through different instructions in CUAD. Our hypothesis of a potential negative transfer is corroborated by our results on LegalBench by categories when we remove the datasets or tasks that overlap between LawInstruct and LegalBench (see Figure 14): We see larger gains compared to the baseline for both



(a) LegalBench (b) MMLU

Figure 6: Difference to the baseline for the XL model across tasks on LegalBench and MMLU. For LegalBench, we excluded tasks with a difference between -10 and 10 for clarity.

the conclusion and the interpretation categories.

4 Ablations

In this section, we perform controlled experiments across the starting checkpoints, data mixtures, instruction styles and amount of instruction tuning data during pretraining. We show additional ablations regarding sampling styles, licenses and crosslingual transfer from multilingual data in Appendix D. Flan-T5 performs best in the studied parameter ranges. Baselines for other models are in Appendix E Table 5.

4.1 Starting Checkpoint

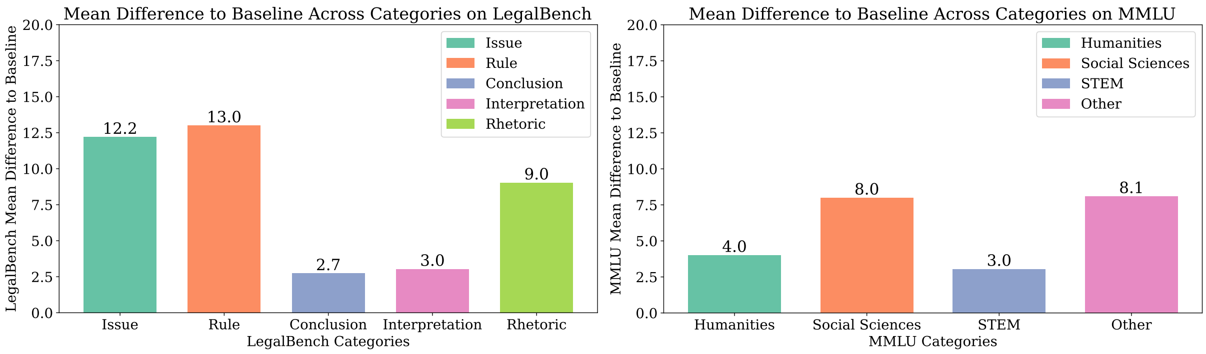
Should you start in-domain instruction tuning from a base model or from an instruction tuned model? ⇒ **Starting from an instruction tuned model is better across sizes except Small.** In Figure 8, we compare instruction tuning from a base T5 and a Flan-T5 model in four different sizes (Small, Base, XL and XXL) (detailed results in Appendix E Table 6). We find that for the larger sizes, the instruction tuned Flan-T5 is a better starting point ($p < 0.001$), leading to higher performance on LegalBench. For the Small size the difference is not statistically significant ($p = 0.058$). We use the Flan-T5 model as a starting point in all experiments unless specified otherwise.

4.2 Data Mixture

What data mixtures should you choose for in-domain instruction tuning? ⇒ **Mixing in general instruction tuning datasets is necessary.** In Figure 9, we compare instruction tuning with three different data mixtures: lawinstruct, flan2 (Chung et al., 2022), and flan2-lawinstruct (where we sample equally from flan2 and lawinstruct) (detailed results in Appendix E Table 7). Interestingly, when only training on lawinstruct, downstream accuracy drops, possibly due to the instructions in our datasets being formulated differently than the original Flan instructions. Training on flan2 and flan2-lawinstruct leads to an aggregate increase of 7.7 points (48.3 to 56) and 10.8 points (48.3 to 59.1) respectively. We use the flan2-lawinstruct mixture in all experiments unless specified otherwise.

4.3 Instruction Style

Are models trained with more diverse instructions better on LegalBench? ⇒ **Results are mixed, overall just using one instruction is probably sufficient.** In Figure 10, we compare the performance of



(a) LegalBench

(b) MMLU

Figure 7: Difference to the baseline for the XL model across categories on LegalBench and MMLU.

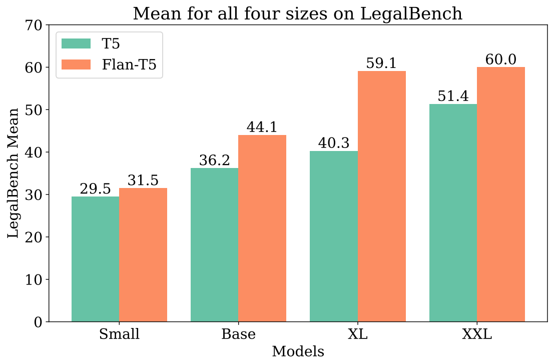


Figure 8: Starting instruction tuning from the Flan-T5 checkpoint improves results across all sizes.

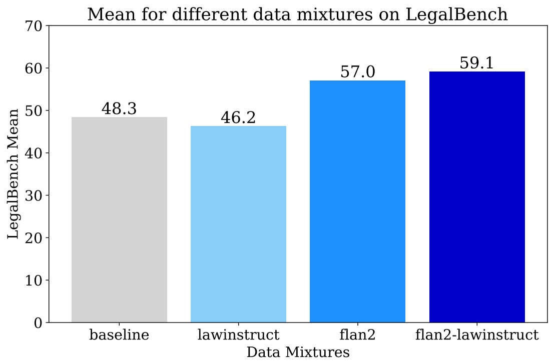


Figure 9: Accuracy of the Flan-T5 XL model on LegalBench using three data mixtures.

training with just one manually written instruction vs. ten paraphrased instructions with GPT-4 from one seed instruction, all else constant (detailed results in Appendix E Table 10). For Flan-T5 (see Table 10), for Small, one instruction is better than ten ($p = 0.035$); for the other sizes we find no difference. For mT5 (see Figure 10b), for Small, one instruction is worse than ten both monolingual ($p = 0.005$) and multilingual ($p = 0.01$) whereas for XL, ten English instructions underperform one English ($p < 0.001$) and ten multilingual ones ($p < 0.001$). In aggregate, differences are small without a consistent trend.

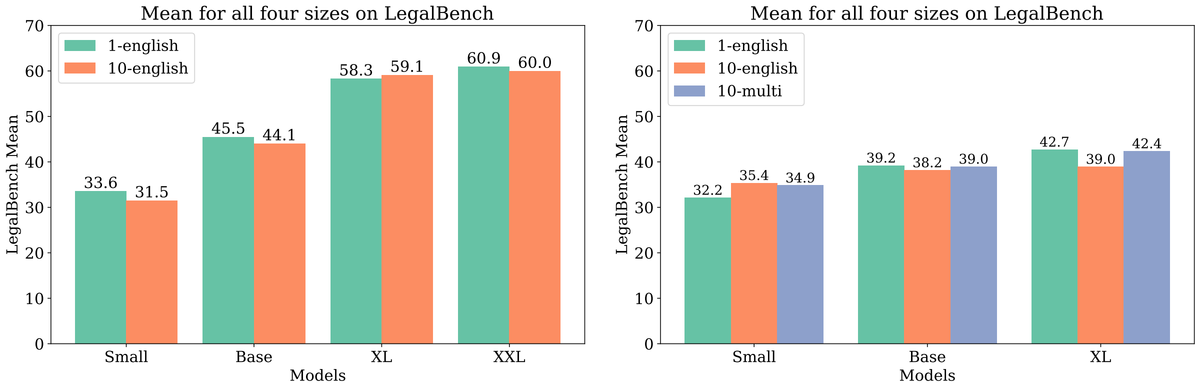
4.4 Amount of Instruction Data During Continued Pretraining

How much instruction tuning data should be mixed in during continued pretraining? $\Rightarrow$ **Continued pretraining seems to be rather robust w.r.t. the amount of instruction tuning samples mixed in.** In Tables 12 to 15, we investigate the benefit of mixing varying amounts of instruction tuning data in during continued pretraining (detailed results in Appendix E Tables 12 to 15). We compare results on LegalBench of instruction tuning runs after 10K to 90K steps of continued pretraining. For the Small model, the benefit of continued pretraining over just instruction tuning is significant (34.9 for just instruction tuning vs. 40 after continued pretraining). Conversely, for the XL model, continued pretraining often underperforms compared to just instruction tuning. For the XXL model, more instruction tuning samples during continued pretraining improve performance, unlike for the Small and XL models. Across sizes, continued pretraining's effectiveness appears robust to the number of instruction tuning samples used.[4]

5 Related Work

Domain-specific pretraining, covering areas such as medicine, law, and science, significantly enhances Language Model performance on related tasks (Beltagy et al., 2019; Gu et al., 2021; Chalkidis et al., 2020). SciBERT (Beltagy et al., 2019), for instance, was pretrained on a mix of computer science and biomedical papers, exemplifying this approach in the scientific domain. Other models like PubMedBERT (Gu et al., 2021) and BioBERT (Lee et al., 2020), specifically pretrained on biomedical datasets, have shown improvements

[4]Mixing instruction tuning data during continued pretraining without more instruction tuning does not improve results.



(a) English

(b) Multilingual

Figure 10: Ablation on the instruction style on English/multilingual flan2-lawinstruct from the Flan-T5/mT5 checkpoint across all sizes.

in medical NLP tasks (Huang et al., 2019).

5.1 Domain-specific Legal Pretraining

In the legal domain, models such as LegalBERT, pretrained on 12 GB of English legal texts, demonstrated notable success in domain-specific challenges (Chalkidis et al., 2020). CaseLaw-BERT capitalized on the English Harvard Law case corpus spanning from 1965 to 2021 (Zheng et al., 2022), while Niklaus and Giofré (2022) pretrained LongFormer models on the Pile-of-Law (Henderson et al., 2022) using the replaced token detection task (Clark et al., 2020) for enhanced performance. Further advancements were made by Chalkidis et al. (2023), who developed new English legal LMs yielding superior results on LexFiles, a compilation of 11 sub-corpora from six English-speaking legal systems encompassing 19B tokens. Additionally, Niklaus et al. (2023b) introduced a vast multilingual legal corpus, training both monolingual and multilingual legal models to achieve state-of-the-art results on LexGLUE (Chalkidis et al., 2022) and LEXTREME (Niklaus et al., 2023a). Models have also been developed for specific jurisdictions, including the Swiss (Rasiah et al., 2023), Italian (Licari and Comandè, 2022), Romanian (Masala et al., 2021), and Spanish (Gutiérrez-Fandiño et al., 2021) legal systems. Despite the prevalence of smaller encoder-based legal-specific LMs, larger generative models in this space remain scarce. This work seeks to bridge that gap.

5.2 Instruction Tuning

Instruction tuning – the process of finetuning auto-regressive pretrained language models on corpora of reciprocal instruction–response pairs – has emerged as a critical step for building responsive models that are useful for many tasks (Ouyang et al., 2022; Chowdhery et al., 2022; Wei et al., 2022b; Sanh et al., 2022). Some go as far as to claim that this training paradigm is the key to imbuing language models with the generalized capability of zero-shot instruction following behavior (Chung et al., 2022). Instruction tuning refers to few-shot or zero-shot adaptation of large language models to new tasks, where the task is described in natural language in the training examples. Following Wei et al. (2022a), it is common to transform existing datasets into instruction datasets by manually composing templates and filling these with specific examples. It is through these domain-specific training procedures that we build and evaluate legal data adaptation in LLMs.

6 Conclusion and Future Work

We curated LawInstruct, the first instruction tuning dataset for the legal domain by aggregating various high-quality annotated datasets and writing instructions for the different tasks. We used LawInstruct to instruction tune T5 based models, creating FLawN-T5 and a new state-of-the-art on LegalBench in all investigated parameter sizes. We openly release LawInstruct on Hugging Face.

In the future, we would like to extend LawInstruct with more high-quality datasets released after our experiments such as Negation Scope Resolution (Christen et al., 2023), or Legal Violation Detection (Bernsohn et al., 2024). Additionally, it would be interesting to investigate overlap between the T5 pretraining dataset C4 and the MultiLegalPile to get a better understanding of the potential benefits of continued pretraining.

Limitations

Our use of template-based instruction creation may restrict the variety of instructions, potentially affecting the model's ability to handle more diverse or novel legal queries effectively. While we already tried to address this by paraphrasing the instructions with GPT-4, the diversity may still be limited. To alleviate this problem, we could create synthetic data either by generating responses from instructions (Wang et al., 2023c) or reversely, by generating instructions to responses (Köksal et al., 2024). It is important to take care to do detailed quality checks since hallucinated content may hurt more than improve, especially in the legal domain. Another way to alleviate this diversity problem is working with legal professionals to identify and annotate new tasks for the legal domain. However, this route is out of reach for many academic efforts due to large salaries of qualified lawyers.

To our surprise, continued pretraining only benefited at the Small model size, but not at larger sizes. Due to our focus on instruction tuning and limited budget, we were not able to study this effect in more detail. In future work, we would like to study the robustness of our findings across model sizes. We hypothesize that methods like mixing in data from the original training set, using smaller learning rates, and adding loss terms to discourage the weights to depart too much from the original model could potentially lead to different conclusions.

References

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. *arXiv preprint*. ArXiv:2210.13448 [cs].

Nikos Bartziokas, Thanassis Mavropoulos, and Constantine Kotropoulos. 2020. Datasets and Performance Metrics for Greek Named Entity Recognition. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020)*, SETN 2020, pages 160–167, New York, NY, USA. Association for Computing Machinery.

Shrutarshi Basu, Nate Foster, James Grimmelmann, Shan Parikh, and Ryan Richardson. 2022. A programming language for future interest. *Yale JL & Tech.*, 24:75.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text. *arXiv preprint*. ArXiv:2402.04335 [cs].

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *European Conference on Information Retrieval*, pages 413–428. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

William Bruno and Dan Roth. 2022. LawngNLI: A Long-Premise Benchmark for In-Domain Generalization from Short to Long Contexts and for Implication-Based Retrieval. *arXiv preprint arXiv:2212.03222*.

CAIL 2022. 2022. CAIL 2022. `https://github.com/china-ai-law-challenge/CAIL2022`.

Pablo Calleja, Patricia Martín Chozas, Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel, Elsa Gómez, and Pascual Boil. 2021. Bilingual dataset for information retrieval and question answering over the spanish workers statute. In *XIX Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA)*.

Case briefs. 2024. Case briefs. `https://www.oyez.org/`.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv:2109.00904 [cs]*. ArXiv: 2109.00904.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021b. Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. *Preprint*, arXiv:2305.07507.

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 *ACL (1)*, pages 4310–4330. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv:2204.02311 [cs]*. ArXiv: 2204.02311.

Ramona Christen, Anastassia Shaitarova, Matthias Stürmer, and Joel Niklaus. 2023. Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents. *arXiv preprint*. ArXiv:2309.08695 [cs].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint*. ArXiv:2210.11416 [cs].

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*. ArXiv: 2003.10555.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. SaulLM-7B: A pioneering Large Language Model for Law. *arXiv preprint*. ArXiv:2403.03883 [cs].

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *arXiv preprint*. ArXiv:2401.01301 [cs].

Ona de Gibert Bonet, Aitor García Pablos, Montse Cuadros, and Maite Melero. 2022. Spanish datasets for sensitive entity detection in the legal domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3751–3760, Marseille, France. European Language Resources Association.

Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and Alexandre Rademaker. 2017. Passing the Brazilian OAB exam: data preparation and some experiments. *Preprint*, arXiv:1712.05128. ArXiv preprint arXiv:1712.05128.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A Corpus for Multilingual Analysis of Online Terms of Service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 1–8, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. 2022. Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8577–8591, Dublin, Ireland. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1).

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].

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish Legalese Language Model and Corpora. *arXiv preprint*. ArXiv:2110.12201 [cs].

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2022. Mining Legal Arguments in Court Decisions. *arXiv preprint*.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. *arXiv preprint*. ArXiv:2207.00220 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021c. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *arXiv preprint*. ArXiv:2103.06268 [cs].

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *NLLP@KDD*, pages 31–38.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction. *arXiv preprint*. ArXiv:2206.05224 [cs].

Elias Jacob de Menezes-Neto and Marco Bruno Miranda Clementino. 2022. Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts. *PLOS ONE*, 17(7):e0272287.

Heewon Jeon. 2021. Legalqa using sentencekobart. https://github.com/haven-jeon/LegalQA.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. GPT-4 Passes the Bar Exam.

Moniba Keymanesh, Micha Elsner, and Srinivasan Sarthasarathy. 2020. Toward domain-guided controllable summarization of privacy policies. In *NLLP@ KDD*, pages 18–24.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2022. Coliee 2022 summary: Methods for legal document retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 51–67. Springer.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. 2022. Validity assessment of legal will statements as natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6047–6056, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2024. LongForm: Effective Instruction Tuning with Reverse Instructions. *arXiv preprint*. ArXiv:2304.08460 [cs].

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting Brazilian Court Decisions. *PeerJ Computer Science*, 8:e904. Publisher: PeerJ Inc.

Law Stack Exchange. 2024. Law stack exchange. https://law.stackexchange.com/.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

LegalQA. 2019. LegalQA. https://github.com/siatnlp/LegalQA.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniele Licari and Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 313–323, Cham. Springer International Publishing.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Laura Manor and Junyi Jessy Li. 2019. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. *arXiv preprint*. ArXiv:2305.16938 [cs].

Emre Mumcuoğlu, Ceyhun E. Öztürk, Haldun M. Ozaktas, and Aykut Koç. 2021. Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing & Management*, 58(5):102684.

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*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? *arXiv preprint*. ArXiv:2211.17135 [cs].

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023a. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. *arXiv preprint*. ArXiv:2301.13126 [cs].

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023b. MultiLegalPile: A 689GB Multilingual Legal Corpus. *arXiv preprint*. ArXiv:2306.02069 [cs].

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*.

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text

Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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].

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv:2110.08207 [cs]*. ArXiv: 2110.08207.

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.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. *arXiv preprint*. ArXiv:2206.10883 [cs].

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. 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*, pages 1048–1064.

Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2020. Supreme Court Database, Version 2020 Release 01.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying Language Learning Paradigms. *arXiv preprint*. ArXiv:2205.05131 [cs].

Nguyen Ha Thanh, Bui Minh Quan, Chau Nguyen, Tung Le, Nguyen Minh Phuong, Dang Tran Binh, Vuong Thi Hai Yen, Teeradaj Racharak, Nguyen Le Minh, Tran Duc Vu, Phan Viet Anh, Nguyen Truong Son, Huy Tien Nguyen, Bhumindr Butr-indr, Peerapon Vateekul, and Prachya Boonkwan. 2021. A summary of the alqac 2021 competition. In *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–5.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Georgios Tziafas, Eugenie de Saint-Phalle, Wietse de Vries, Clara Egger, and Tommaso Caselli. 2021. A multilingual approach to identify and classify exceptional measures against covid-19. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 46–62. Dataset URL: https://tinyurl.com/ycysvtbm.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2021. Design and Implementation of German Legal Decision Corpora:. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pages 515–521, Online Streaming, — Select a Country —. SCITEPRESS - Science and Technology Publications.

Maarten Peter Vink, Luuk Van Der Baaren, Rainer Bauböck, Jelena DZANKIC, Iseult HONOHAN, and Bronwen MANBY. 2021. Globalcit citizenship law dataset.

Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. *ASAIL@ ICAIL*, 2385.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. page 30.

Steven H. Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023a. MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding. *arXiv preprint*. ArXiv:2301.00876 [cs].

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *Preprint*, arXiv:2306.04751.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. *arXiv preprint*. ArXiv:2204.07705 [cs].

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022b. Finetuned Language Models Are Zero-Shot Learners. *arXiv preprint*. ArXiv:2109.01652 [cs].

Benjamin Weiser. 2023. Here's what happens when your lawyer uses chatgpt. *New York Times*.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*. ArXiv: 1807.02478.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Heng Wang, Jianfeng Xu, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*. ArXiv: 2010.11934.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, Yu-Cheng Zhou, and Jia-Rui Lin. 2022. Pretrained Domain-Specific Language Model for Natural Language Processing Tasks in the AEC Domain. *Comput. Ind.*, 142(C). Place: NLD Publisher: Elsevier Science Publishers B. V.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jecqa: A legal-domain question answering dataset. In *Proceedings of AAAI*.

A Use of AI Assistants

We used ChatGPT 3.5 and 4 for shortening texts and editing LaTeX more efficiently.

B Detailed Dataset Description

Figure 11 shows the LawInstruct task type and jurisdiction composition by dataset. Table 2 lists the dataset (and sources), license, language, jurisdiction, task type, subtask, and number of examples for each dataset included in LawInstruct.

C Detailed Experimental Setup

C.1 Inexplicable Behaviour at the XXL Size

We spent considerable effort, including joint debugging with the authors of LegalBench, to reproduce

Table 2: Overview of the LawInstruct datasets. The 24 EU langs are bg, cs, da, de, el, en, es, et, fi, fr, ga, hu, it, lt, lv, mt, nl, pt, ro, sv, sk. Abbreviations: Terms of Service (ToS)

Dataset	License	Languages	Jurisdiction	Tasks	Subtask	Examples
Benchmark for Understanding Indian Legal Documents (BUILD) (Kalamkar et al., 2022)	Unknown	en	India	Text classification	Rhetorical role	28,986
Brazilian Bar Exam (Delfino et al., 2017)	Unknown	pt	Brazil	Question answering	Bar exam questions	2,130
Brazilian Court Decisions (Lage-Freitas et al., 2022)	Unknown	pt	Brazil	Text classification	Judgment	3,234
Brazilian Court Decisions (Lage-Freitas et al., 2022)	Unknown	pt	Brazil	Text classification	Decision Unanimity	1,715
BrCAD5 (Jacob de Menezes-Neto and Clementino, 2022)	CC BY-NC-SA 4.0	pt	Brazil	Multiple choice	Judgment	1,225,922
BrCAD5 (Jacob de Menezes-Neto and Clementino, 2022)	CC BY-NC-SA 4.0	pt	Brazil	Text classification	Judgment	612,961
BrCAD5 (Jacob de Menezes-Neto and Clementino, 2022)	CC BY-NC-SA 4.0	pt	Brazil	Text classification	Area of law	612,961
BrCAD5 (Jacob de Menezes-Neto and Clementino, 2022)	CC BY-NC-SA 4.0	pt	Brazil	Text classification	Topic	1,838,883
BVADecisions (Walker et al., 2019)	MIT	en	USA	Text classification	Rhetorical role	8,818
BVADecisions (Walker et al., 2019)	MIT	en	USA	Question answering	Relevant rules	2
CAIL 2019 (Xiao et al., 2019)	Unknown	zh	China	Question answering	Chinese legal case questions	39,333
CAIL 2022 (CAIL 2022)	Unknown	zh	China	Text classification	Charge/crime	10,448
CAIL 2022 (CAIL 2022)	Unknown	zh	China	Argument & counter-argument	5,224	
CAIL 2022 (CAIL 2022)	Unknown	zh	China	Question answering	Response to argument	5,224
Case Briefs (Case briefs)	CC BY-NC	en	USA	Question answering	Legal analysis of facts	2,619
CaseHOLD (Zheng et al., 2021)	CC-BY	en	USA	Multiple choice	Legal holding statements	45,000
Change My View (Tan et al., 2016)	Unknown	en	N/A	Argument & counter-argument	3,456	
COLIEE (Kim et al., 2022)	Academic use only	en, jp	Canada/Japan	Question generation	Entailed question	1,774
COLIEE (Kim et al., 2022)	Academic use only	en, jp	Canada/Japan	Natural language inference	Passage entailment	125,954
COLIEE (Kim et al., 2022)	Academic use only	en, jp	Canada/Japan	Question answering	Relevant legal rule	1,774
ContractNLI (Koreeda and Manning, 2021)	CC BY-NC	en	USA	Natural language inference	Premise hypothesis entailment	14,010
COVID-19 Emergency Measures (EXCEPTIUS) (Tziafas et al., 2021)	Unknown	en, fr, hu, it, nb, nl, pl	EU	Text classification	Measure type	3,312
European Court of Human Rights (ECtHR) (Chalkidis et al., 2021b)	CC BY-NC-SA 4.0	en	EU	Text classification (multi-label)	Violated article	9,000
European Court of Human Rights (ECtHR) (Chalkidis et al., 2021b)	CC BY-NC-SA 4.0	en	EU	Text classification (multi-label)	Allegedly violated article	9,000
EOIR (Henderson et al., 2022)	CC BY-NC-SA 4.0	en	USA	Text classification	Pseudonymity	8,089
EURLEX (Chalkidis et al., 2019)	CC BY-SA 4.0	en	EU	Text classification	EuroVoc core concepts	55,000
EUR-Lex-Sum (Aumiller et al., 2022)	CC BY 4.0	24 EU langs	EU	Summarization	EU Legal Acts	22,989
German Argument Mining (Urchs et al., 2021)	CC BY 4.0	de	Germany	Text classification	Argumentative function	19,271
German Rental Agreements (Steinberger et al., 2006)	Unknown	de	Germany	Text classification	Semantic type	3,292
Greek Legal Code (Papaloukas et al., 2021)	CC BY 4.0	el	Greece	Text classification	Volume (coarse thematic topic)	28,536
Greek Legal Code (Papaloukas et al., 2021)	CC BY 4.0	el	Greece	Text classification	Chapter (intermediate thematic topic)	28,536
Greek Legal Code (Papaloukas et al., 2021)	CC BY 4.0	el	Greece	Text classification	Subject (fine-grain thematic topic)	28,536
Greek Legal NER (elNER) (Bartziokas et al., 2020)	CC BY-NC-SA 4.0	el	Greece	Named entity recognition	Greek legal entities	17,699
ILDC (Malik et al., 2021)	CC BY-NC	en	India	Text classification	Judgment	37,387
International Citizenship Law (Vink et al., 2021)	CC BY 4.0	en	International	Question answering	Citizenship acquisition	6,460
International Citizenship Law (Vink et al., 2021)	CC BY 4.0	en	International	Question answering	Citizenship loss	2,850
JEC-QA (Zhong et al., 2020)	CC BY-NC-ND	zh	China	Multiple choice	National Judicial Examination of China	21,072
Korean Legal QA (Jeon, 2021)	Academic use only	ko	South Korea	Question answering	Relevant law	1,830
LawngNLI (Bruno and Roth, 2022)	MIT	en	USA	Natural language inference	Premise hypothesis entailment	1,142,304
LBOX OPEN (Hwang et al., 2022)	CC BY-NC	ko	South Korea	Text classification	Judgment	12,142
LBOX OPEN (Hwang et al., 2022)	CC BY-NC	ko	South Korea	Text classification	Relevant statutes	13,317
LEDGAR (Tuggener et al., 2020)	CC BY-NC	en	USA	Text classification	Contract provision category	60,000
Legal Case Document Summarization (Shukla et al., 2022; Bhattacharya et al., 2019)	CC BY-SA	en	India	Summarization	Indian Supreme Court	7,080
Legal Case Summarization (Shukla et al., 2022; Bhattacharya et al., 2019)	CC BY-SA	en	UK	Summarization	UK Supreme Court	693
LegalNERo (Pais et al., 2021)	CC0 1.0	ro	Romania	Named entity recognition	Romanian legal entities	7,552
LegalQA (LegalQA)	Unknown	zh	China	Question answering	Legal advice	21,946
LeNER-Br (Luz de Araujo et al., 2018)	Unknown	pt	Brazil	Named entity recognition	Brazilian legal entities	7,828
Littleton (Basu et al., 2022)	MIT	en	USA	Question answering	Relevant future interests	131
Littleton (Basu et al., 2022)	MIT	en	USA	Question answering	Event graph	143
MAPA (de Gibert Bonet et al., 2022)	CC BY-NC 4.0	24 EU langs	EU	Named entity recognition	Coarse-grained	27,823
MAPA (de Gibert Bonet et al., 2022)	CC BY-NC 4.0	24 EU langs	EU	Named entity recognition	Fine-grained	27,823
MAUD (Wang et al., 2023a)	CC BY	en	USA	Multiple choice	Merger agreement questions	10,751
MAUD (Wang et al., 2023a)	CC BY	en	USA	Text classification	Deal point category	25,827
MAUD (Wang et al., 2023a)	CC BY	en	USA	Text classification	Question type	25,827
MAUD (Wang et al., 2023a)	CC BY	en	USA	Text classification	Text type	25,827
Mining Legal Arguments (Habernal et al., 2022)	Apache-2.0	en	EU	Named entity recognition	Actors	31,852
Mining Legal Arguments (Habernal et al., 2022)	Apache-2.0	en	EU	Named entity recognition	Argument type	31,852
MultiEURLEX (Chalkidis et al., 2021a)	CC BY-SA	24 EU langs	EU	Text classification (multi-label)	EuroVoc taxonomy (coarse level)	1,265,000
MultiEURLEX (Chalkidis et al., 2021a)	CC BY-SA	24 EU langs	EU	Text classification (multi-label)	EuroVoc taxonomy (intermediate level)	911,798
MultiEURLEX (Chalkidis et al., 2021a)	CC BY-SA	24 EU langs	EU	Text classification (multi-label)	EuroVoc taxonomy (fine-grain level)	1,265,000
Multi-LexSum (Shen et al., 2022)	ODC-By	en	USA	Summarization	Long to short	2,210
Multi-LexSum (Shen et al., 2022)	ODC-By	en	USA	Summarization	Long to tiny	1,130
Multi-LexSum (Shen et al., 2022)	ODC-By	en	USA	Summarization	Short to tiny	1,129
Natural Instructions (BillSum) (Kornilova and Eidelman, 2019)	CC0 1.0	en	USA	Summarization	U.S Congressional and California state bills	25,200
Natural Instructions (CAIL 2018) (Xiao et al., 2018)	Unknown	zh	China	Question answering	Judgment	5,988
Natural Instructions (CaseHOLD) (Zheng et al., 2021)	CC-BY	en	USA	Multiple choice	Correct answer	5,988
Natural Instructions (CaseHOLD) (Zheng et al., 2021)	CC-BY	en	USA	Multiple choice	Incorrect answer	5,988
Natural Instructions (CUAD) (Hendrycks et al., 2021c)	CC BY 4.0	en	Question answering	Information relevant for contract review	2,442	
Natural Instructions (CUAD) (Hendrycks et al., 2021c)	CC BY 4.0	en	USA	Question generation	Questions relevant for contract review	2,442
Natural Instructions (EURLEX) (Chalkidis et al., 2019)	CC BY-SA 4.0	en	EU	Text classification	Regulation, decisions, or directive	5,850
Natural Instructions (EURLEX) (Aumiller et al., 2022)	CC BY-SA 4.0	en	EU	Summarization	EU Legal Acts	3,900
Natural Instructions (OPP-115) (Wilson et al., 2016)	CC BY-NC	en	USA	Question answering	Type of information used by website	18,480
Natural Instructions (OPP-115) (Wilson et al., 2016)	CC BY-NC	en	USA	Question answering	Purpose of privacy policy	18,474
Natural Instructions (Overruling) (Zheng et al., 2021)	Unknown	en	USA	Text classification	Sentence is overruling	14,370
OLC Memos (Henderson et al., 2022)	CC BY-NC	en	USA	Question answering	Write a legal research memo	1,038
Online ToS (Drawzeski et al., 2021)	CC BY-NC 2.5	de, en, it, pt	Unknown	Text classification	Clause topic	19,942
Online ToS (Drawzeski et al., 2021)	CC BY-NC 2.5	de, en, it, pt	Unknown	Text classification	Unfair contractual term type	2,074
Plain English Contracts Summarization (Manor and Li, 2019)	Unknown	en	USA	Summarization	Software licenses, ToS	446
PrivacyQA (Ravichander et al., 2019)	MIT	en	Unknown	Question answering	Contents of privacy policies	185,200
PrivacySummarization (Keymanesh et al., 2020)	MIT	en	USA	Summarization	Privacy policies, ToS, and cookie policies	5,751
RedditLegalQA (Henderson et al., 2022)	CC BY 4.0	en	Unknown	Question answering	Legal advice from r/legaladvice	192,953
Sara (Holzenberger et al., 2020)	Unknown	en	USA	Natural language entailment	Fact entailment	176
Sara (Holzenberger et al., 2020)	Unknown	en	USA	Question answering	Tax liability	160
SaraProlog (Holzenberger et al., 2020)	Unknown	en	USA	Question answering	Fact pattern to prolog code	376
SaraProlog (Holzenberger et al., 2020)	Unknown	en	USA	Question answering	Tax statute to prolog code	9
Short Answer Feedback (Filighera et al., 2022)	CC BY 4.0	de	Germany	Question answering	Answer question about German law	1,596
Short Answer Feedback (Filighera et al., 2022)	CC BY 4.0	de	Germany	Question answering	Feedback rating for answer	1,596
Spanish Labor Law (Calleja et al., 2021)	CC BY 4.0	es	Spain	Extractive question answering	Answer question about Spanish labor law	111
StackExchange Questions (Law) (Law Stack Exchange)	CC BY-SA	en	Unknown	Question answering	Online legal forum	10,158
The Supreme Court Database (Spaeth et al., 2020)	CC BY-NC 3.0	en	USA	Text classification	Issue areas	5,000
Swiss Federal Supreme Court (Rasiah et al., 2023)	CC BY 4.0	de, fr	Text generation	Case considerations sections (lower court)	26	
Swiss Courts (Rasiah et al., 2023)	CC BY 4.0	de, fr, it	Switzerland	Text generation	Case considerations sections (same court)	234,313
Swiss Federal Supreme Court (Rasiah et al., 2023)	CC BY 4.0	de, fr, it	Switzerland	Text classification	Case criticality (based on citations)	91,075
Swiss Courts (Rasiah et al., 2023; Niklaus et al., 2021)	CC BY 4.0	de, fr, it, en	Switzerland	Multiple choice	Judgment	477,636
Swiss Courts (Rasiah et al., 2023; Niklaus et al., 2021)	CC BY 4.0	de, fr, it	Switzerland	Text classification	Judgment	385,719
Swiss Courts (Rasiah et al., 2023; Niklaus et al., 2021)	CC BY 4.0	de, fr, it, en	Switzerland	Text classification	Area of law	18,162
Swiss Courts (Rasiah et al., 2023; Niklaus et al., 2021)	CC BY 4.0	de, fr, it, en	Switzerland	Text classification	Subarea of law	18,162
Swiss Federal Supreme Court (Leading Decisions) (Rasiah et al., 2023)	CC BY 4.0	de, en, fr, it	Switzerland	Text classification	Location (canton, region)	42,342
Swiss Legislation (Rasiah et al., 2023)	CC BY 4.0	de, fr, it, rm	Switzerland	Text classification	Abbreviation	11,045
Swiss Legislation (Rasiah et al., 2023)	CC BY 4.0	de, en, fr, it, rm	Switzerland	Text classification	Canton	35,698
Swiss Legislation (Rasiah et al., 2023)	CC BY 4.0	de, en, fr, it, rm	Switzerland	Text classification	Short description	3,747
Swiss Legislation (Rasiah et al., 2023)	CC BY 4.0	de, en, fr, it, rm	Switzerland	Text classification	Title	35,359
Thai Supreme Court Cases (TSCC)(Thanh et al., 2021)	Academic use only	th	Thailand	Question answering	Relevant legal articles (Thai Criminal Code)	2,883
Turkish Constitutional Court (Mumcuoğlu et al., 2021)	CC BY 4.0	tr	Turkey	Multiple choice	Judgment	1,804
Turkish Constitutional Court (Mumcuoğlu et al., 2021)	CC BY 4.0	tr	Turkey	Text classification	Judgment	902
Unfair ToS (Lippi et al., 2019)	Unknown	en	USA	Text classification (multi-label)	Unfair contractual term type	5,532
U.S Class Actions (Semo et al., 2022)	GPL-3.0	en	USA	Text classification	Judgment	3,000
Valid Wills (Kwak et al., 2022)	Unknown	en	USA	Text classification	Statement supported by law/condition	1,512

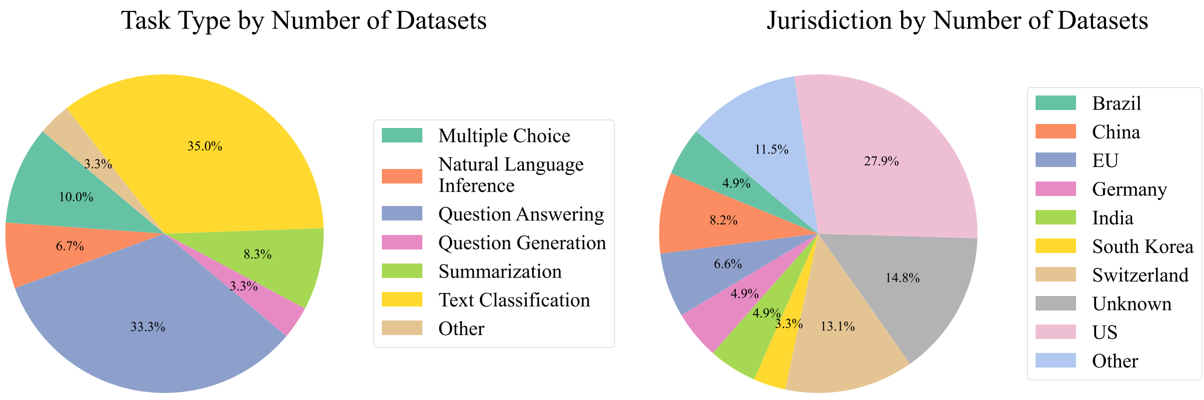


Figure 11: Jurisdiction and task type by datasets.

their results. We double checked that the prompts, decoding hyperparameters and general setup are consistent. We conjecture, that the conversion of the Flan-T5 weights as done by Hugging Face on their hub leads to different behavior when running the models with T5X on TPUs (our setup) vs running them with Hugging Face transformers and PyTorch on NVIDIA GPUs (original LegalBench setup)[5].

The XXL mT5 model did not train stably in the continued pretraining phase despite heavy hyperparameter tuning.

C.2 Evaluation

We excluded any legal tasks occurring in MMLU from LawInstruct. However, there is some overlap regarding the tasks included in LawInstruct and in LegalBench because high-quality legal tasks are rare. To control for these overlapping tasks, we evaluate on two versions of LegalBench holding out tasks by the datasets or tasks occurring in LawInstruct respectively.

C.2.1 LegalBench Dataset Held Out

If the source dataset of the LegalBench task occurs in LawInstruct, we remove it from the evaluation. Below, we list which tasks are overlapping. Overall 100 tasks are held out (see Table 3 for the complete list), so 61 tasks are remaining for LegalBench evaluation.

C.2.2 LegalBench Task Held Out

We additionally catalog instructions which train the LLM for a task captured in LegalBench. It is not necessary that the instruction-response pair in LawInstruct contain data from LegalBench, just that they are about similar legal tasks (e.g., classifying choice-of-forum provisions). In Table 4, we list which tasks are overlapping. Overall 64 tasks are held out, so 97 tasks are remaining for LegalBench evaluation.

[5]Similar issues are mentioned in this issue: https://github.com/PiotrNawrot/nanoT5/issues/25

D Additional Ablations

D.1 Sampling Style

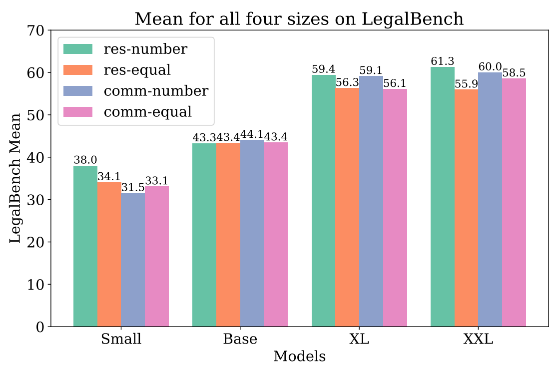


Figure 12: Ablation on sampling style and license on English flan2-lawinstruct from the Flan-T5 checkpoint across sizes. Abbreviations: *res*: licensed for research use (all datasets), *comm*: commercially friendly licensed, *number*: sampling by the number of examples per dataset, *equal*: equally sampling from each dataset

Should we sample each dataset equally or rather by the number of examples? ⇒ **Sampling by the number of examples generally leads to better performance.** In Figure 12, we compare the performance of two sampling styles (equal sampling of each dataset and sampling by the number of examples) across both the research and commercial licensed dataset (detailed results in Appendix E

Table 3: LegalBench Dataset Held Out

Dataset	LawInstruct	LegalBench
ContractNLI	ContractNLI-contract_nli	contract_nli_*
CUAD	NaturalInstructionsLegal-cuad_answer_generation, NaturalInstructionsLegal-cuad_question_generation	cuad_*
GLOBALCIT Citizenship Law Dataset	InternationalCitizenshipLawQuestions-international_citizenship_law_questions_mode_acq, InternationalCitizenshipLawQuestions-international_citizenship_law_questions_mode_loss	international_citizenship_questions
MAUD	MAUD-answer, MAUD-category, MAUD-question, MAUD-text_type	maud_*
OPP-115 (Online Privacy Policies, set of 115) Corpus	NaturalInstructionsLegal-online_privacy_policy_text_information_type_generation, NaturalInstructionsLegal-online_privacy_policy_text_purpose_answer_generation	opp_115_*
Overruling	NaturalInstructionsLegal-overruling_legal_classification	overruling
PrivacyQA	PrivacyQA-privacy_qa	privacy_policy_qa
Note: The LegalBench privacy_policy_entailment Source field is currently incorrectly linked to this dataset (PrivacyQA), but is derived from a different dataset (APP-350 Corpus).		
StAtutory Reasoning Assessment (SARA)	Sara-sara_entailment, Sara-sara_tax_liability, SaraProlog-sara_prolog_facts, SaraProlog-sara_prolog_statute	sara_* (built off of SARA v2)
Unfair Terms of Service	LexGLUE-unfair_tos, LEXTREME-online_terms_of_service_clause_topics (multilingual version), LEXTREME-online_terms_of_service_unfairness_levels (multilingual version)	unfair_tos

Table 4: LegalBench Task Held Out

Task	LawInstruct	LegalBench
Rhetorical Role Labeling	bva_decisions_label, indian_text_segmentation, german_argument_mining	function_of_decision_section, oral_argument_question_purpose
Civil Procedure Questions	civipro_questions_generate_*	diversity_*, personal_jurisdiction
Legal Entailment	coliee_task3_passage_entailment, contract_nli, lawng_nli_entailment	contract_nli_*
Contractual Clause Classification	unfair_tos, german_rental_agreements	cuad_*, jcrew_blocker, unfair_tos, contract_qa

Table 8). For the XL and XXL sizes, sampling by the number of examples is better than equal weight for datasets for both the research and commercial datasets, although not always statistically significant (XL res $p = 0.049$, XL comm $p = 0.052$, XXL res $p < 0.001$, XXL comm $p = 0.31$). For the Small size, sampling by the number of examples is better for the research dataset ($p < 0.001$) but not for the commercial dataset ($p = 0.099$), while there is no difference for the Base size. By default, we sample by the number of examples in all following experiments unless specified otherwise.

D.2 License of Instruction Tuning Datasets

Do we need data licensed non-commercially for good performance? $\Rightarrow$ **The commercially licensed data seems to be enough for the larger models.** In Figure 12, we compare the performance of two differently licensed datasets (research and commercial licenses) across both sampling each dataset equally and by the number of examples (detailed results in Appendix E Table 8). There are fewer datasets available with more permissive licenses allowing for commercial use than for research use (see Table 2 for details on licenses). Except for Small size ($p < 0.001$), using more diverse data available only for research shows no significant benefit. By default, we use the commercially licensed dataset in all subsequent experiments unless specified.

D.3 Crosslingual Transfer from Multilingual Data

Is there crosslingual transfer from multilingual data? $\Rightarrow$ **On the English LegalBench, we do not see any crosslingual transfer.** In Figure 13, we compare the performance of the complete multilingual instruction dataset and the English subset across two differently licensed datasets (research and commercial licenses). We see no statistically significant difference between the multilingual training and the English training. We also see no difference between the differently licensed

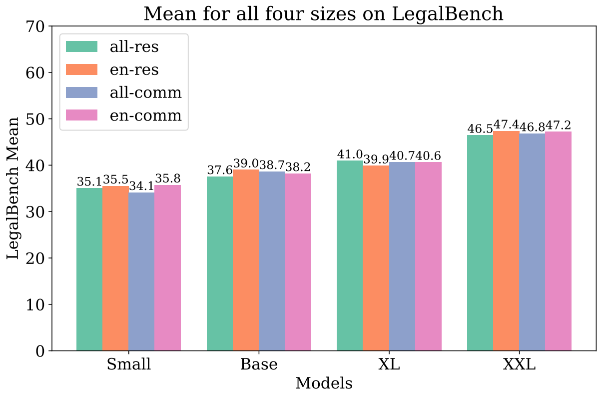


Figure 13: Ablation on the language and license on flan2-lawinstruct from the mT5 checkpoint across all sizes, sampling by the number of examples.

datasets. This means that just training on the commercial subset is enough. We show detailed results on individual LegalBench categories in Appendix E Table 9. Per default we use the English dataset in all following experiments unless specified otherwise.

E Detailed Results

Table 5: Baseline results on LegalBench.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Flan-T5 XXL (ours)	36.1	18.8	25.2	35.1	41.1	31.3
Flan-T5-XXL (Guha et al., 2023)	66.0	36.0	63.3	64.4	70.7	60.1
LLaMA-2-13B (Guha et al., 2023)	50.2	37.7	59.3	50.9	54.9	50.6
OPT-13B (Guha et al., 2023)	52.9	28.4	45.0	45.1	43.2	42.9
Vicuna-13B-16k (Guha et al., 2023)	34.3	29.4	34.9	40.0	30.1	33.7
WizardLM-13B (Guha et al., 2023)	24.1	38.0	62.6	50.9	59.8	47.1
Flan-T5 XL (ours)	53.5	32.1	46.8	58.7	59.6	50.1
Flan-T5-XL (Guha et al., 2023)	56.8	31.7	52.1	51.4	67.4	51.9
BLOOM-3B (Guha et al., 2023)	47.4	20.6	45.0	45.0	36.4	38.9
Incite-3B-Instruct (Guha et al., 2023)	51.1	26.9	47.4	49.6	40.2	43.0
OPT-2.7B (Guha et al., 2023)	53.7	22.2	46.0	44.4	39.8	41.2
Flan-T5 Base (ours)	44.7	18.0	20.9	28.9	37.0	29.9
Flan-T5 Small (ours)	0.3	30.4	39.8	28.2	27.7	25.3

Table 6: The T5 and Flan-T5 models finetuned on flan2-lawinstruct in four sizes.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Small T5	45.5 ± 13.2	25.0 ± 28.9	25.6 ± 27.4	18.6 ± 23.6	32.9 ± 26.8	29.5 ± 10.3
Small Flan-T5	25.0 ± 22.0	38.1 ± 25.4	33.1 ± 24.4	20.6 ± 26.4	40.7 ± 19.5	31.5 ± 8.5
Base T5	49.8 ± 0.7	38.1 ± 25.4	34.0 ± 23.3	21.3 ± 22.8	38.0 ± 19.4	36.2 ± 10.2
Base Flan-T5	50.3 ± 2.4	38.8 ± 25.9	34.0 ± 22.4	43.0 ± 21.1	54.1 ± 13.0	44.1 ± 8.2
XL T5	47.8 ± 12.5	37.5 ± 25.0	38.2 ± 15.5	28.6 ± 25.1	49.4 ± 8.1	40.3 ± 8.5
XL Flan-T5	65.7 ± 15.2	45.1 ± 30.3	49.0 ± 23.5	56.8 ± 18.8	79.0 ± 11.4	59.1 ± 13.6
XXL T5	52.7 ± 6.8	38.5 ± 25.7	50.0 ± 22.8	44.9 ± 25.2	70.7 ± 20.5	51.4 ± 12.1
XXL Flan-T5	55.2 ± 23.7	46.3 ± 31.6	56.1 ± 29.1	57.7 ± 19.8	84.6 ± 9.6	60.0 ± 14.4

Table 7: The Flan-T5 models finetuned on three different data mixtures.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Small baseline	0.3 ± 0.7	30.4 ± 20.3	23.8 ± 25.0	16.9 ± 21.1	32.8 ± 21.4	20.8 ± 13.0
Small lawinstruct	0.0 ± 0.1	15.9 ± 23.9	10.7 ± 22.7	10.5 ± 19.8	18.6 ± 25.7	11.1 ± 7.1
Small flan2	28.2 ± 22.4	37.8 ± 25.3	35.1 ± 24.2	22.6 ± 23.3	40.5 ± 19.4	32.8 ± 7.3
Small flan2-lawinstruct	25.0 ± 22.0	38.1 ± 25.4	33.1 ± 24.4	20.6 ± 26.4	40.7 ± 19.5	31.5 ± 8.5
Base baseline	44.7 ± 12.4	18.0 ± 23.6	36.0 ± 23.8	15.6 ± 19.9	42.7 ± 19.8	31.4 ± 13.8
Base lawinstruct	14.6 ± 14.7	22.3 ± 26.3	30.2 ± 22.6	19.7 ± 26.0	17.8 ± 27.4	20.9 ± 5.9
Base flan2	47.2 ± 4.3	37.6 ± 25.0	28.6 ± 23.4	32.5 ± 21.9	54.4 ± 16.3	40.0 ± 10.6
Base flan2-lawinstruct	50.3 ± 2.4	38.8 ± 25.9	34.0 ± 22.4	43.0 ± 21.1	54.1 ± 13.0	44.1 ± 8.2
XL baseline	53.5 ± 6.0	32.1 ± 24.6	38.2 ± 22.4	49.8 ± 22.6	68.1 ± 20.1	48.3 ± 14.0
XL lawinstruct	54.5 ± 7.7	30.2 ± 35.1	42.9 ± 20.8	39.8 ± 30.8	63.7 ± 14.1	46.2 ± 13.1
XL flan2	65.5 ± 14.6	40.6 ± 27.7	52.0 ± 25.6	53.0 ± 21.9	74.0 ± 20.8	57.0 ± 13.0
XL flan2-lawinstruct	65.7 ± 15.2	45.1 ± 30.3	49.0 ± 23.5	56.8 ± 18.8	79.0 ± 11.4	59.1 ± 13.6
XXL baseline	36.1 ± 21.5	18.8 ± 24.6	39.4 ± 32.1	25.7 ± 24.2	47.6 ± 14.0	33.5 ± 11.4
XXL lawinstruct	54.1 ± 7.2	37.7 ± 27.2	53.2 ± 32.6	46.7 ± 25.0	73.7 ± 15.1	53.1 ± 13.3
XXL flan2	64.0 ± 12.6	44.7 ± 31.4	56.4 ± 27.7	55.5 ± 20.2	81.3 ± 9.7	60.4 ± 13.6
XXL flan2-lawinstruct	55.2 ± 23.7	46.3 ± 31.6	56.1 ± 29.1	57.7 ± 19.8	84.6 ± 9.6	60.0 ± 14.4

Table 8: Flan-T5 models finetuned on four different licence-sampling style configurations.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Small res-number	50.3 ± 1.3	38.2 ± 25.5	34.9 ± 25.6	21.3 ± 26.6	45.3 ± 22.0	38.0 ± 11.1
Small res-equal	34.9 ± 21.2	37.5 ± 25.0	33.0 ± 25.3	21.1 ± 25.1	43.8 ± 19.2	34.1 ± 8.3
Small comm-number	25.0 ± 22.0	38.1 ± 25.4	33.1 ± 24.4	20.6 ± 26.4	40.7 ± 19.5	31.5 ± 8.5
Small comm-equal	31.6 ± 25.1	37.2 ± 24.8	33.6 ± 22.8	20.2 ± 24.0	42.6 ± 21.3	33.1 ± 8.3
Base res-number	49.8 ± 3.2	38.1 ± 25.4	36.0 ± 23.8	42.8 ± 21.3	49.5 ± 12.1	43.3 ± 6.3
Base res-equal	48.9 ± 3.8	39.4 ± 26.3	38.4 ± 25.6	36.6 ± 19.8	53.4 ± 18.3	43.4 ± 7.4
Base comm-number	50.3 ± 2.4	38.8 ± 25.9	34.0 ± 22.4	43.0 ± 21.1	54.1 ± 13.0	44.1 ± 8.2
Base comm-equal	49.2 ± 2.9	38.5 ± 25.7	36.4 ± 20.3	40.5 ± 19.8	52.6 ± 13.3	43.4 ± 7.1
XL res-number	59.9 ± 10.4	44.2 ± 29.8	53.5 ± 28.0	57.1 ± 20.2	82.4 ± 11.1	59.4 ± 14.2
XL res-equal	58.2 ± 8.4	42.3 ± 28.7	46.6 ± 16.8	55.4 ± 19.3	79.0 ± 11.9	56.3 ± 14.3
XL comm-number	65.7 ± 15.2	45.1 ± 30.3	49.0 ± 23.5	56.8 ± 18.8	79.0 ± 11.4	59.1 ± 13.6
XL comm-equal	59.3 ± 10.4	40.6 ± 27.2	47.7 ± 20.7	54.1 ± 20.0	78.7 ± 11.9	56.1 ± 14.4
XXL res-number	62.9 ± 12.3	46.9 ± 31.7	57.6 ± 30.2	56.7 ± 21.5	82.3 ± 9.3	61.3 ± 13.1
XXL res-equal	54.9 ± 6.3	43.3 ± 30.1	55.5 ± 27.3	55.4 ± 19.2	70.5 ± 11.6	55.9 ± 9.6
XXL comm-number	55.2 ± 23.7	46.3 ± 31.6	56.1 ± 29.1	57.7 ± 19.8	84.6 ± 9.6	60.0 ± 14.4
XXL comm-equal	59.5 ± 13.1	45.7 ± 30.0	54.8 ± 27.6	55.4 ± 19.6	77.2 ± 12.3	58.5 ± 11.6

Table 9: Flan-T5 models finetuned on four different language-license configurations.

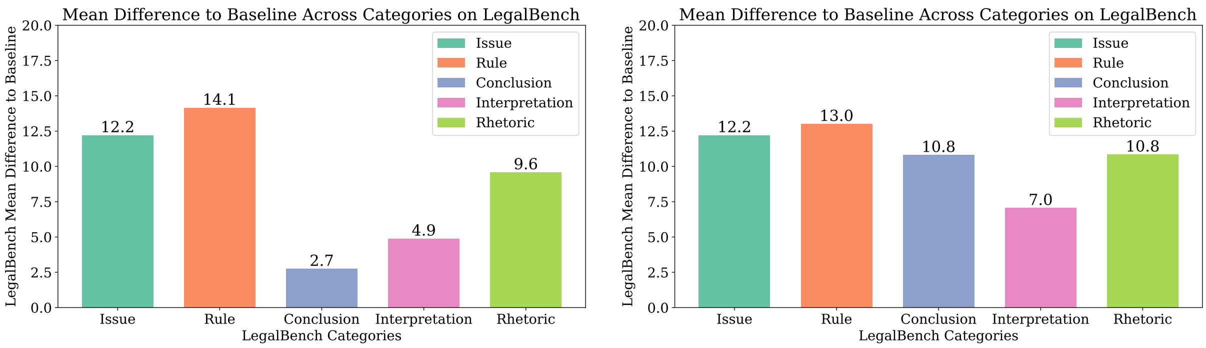
LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Small all-res	46.8 ± 12.2	38.2 ± 24.2	33.8 ± 22.8	20.1 ± 22.0	36.5 ± 21.1	35.1 ± 9.7
Small en-res	50.7 ± 5.9	37.4 ± 24.9	34.0 ± 23.0	18.5 ± 22.9	37.1 ± 23.0	35.5 ± 11.5
Small all-comm	49.7 ± 2.1	38.0 ± 24.1	34.0 ± 23.0	13.2 ± 19.9	35.8 ± 21.8	34.1 ± 13.2
Small en-comm	49.1 ± 13.3	37.5 ± 25.0	34.4 ± 23.3	19.7 ± 23.2	38.2 ± 24.2	35.8 ± 10.6
Base all-res	51.7 ± 4.4	38.7 ± 26.1	33.6 ± 22.7	22.0 ± 23.6	41.8 ± 18.5	37.6 ± 10.9
Base en-res	51.8 ± 5.5	37.5 ± 25.0	37.1 ± 16.5	20.7 ± 22.7	48.0 ± 18.1	39.0 ± 12.1
Base all-comm	51.8 ± 5.2	38.0 ± 25.4	34.3 ± 22.9	23.7 ± 24.7	45.5 ± 12.6	38.7 ± 10.7
Base en-comm	52.0 ± 3.7	37.5 ± 25.0	33.2 ± 22.7	21.9 ± 21.8	46.5 ± 21.2	38.2 ± 11.7
XL all-res	49.9 ± 0.9	37.5 ± 25.0	36.9 ± 18.1	28.3 ± 22.9	52.2 ± 10.7	41.0 ± 9.9
XL en-res	49.9 ± 0.3	37.5 ± 25.0	36.6 ± 18.4	24.8 ± 25.9	50.5 ± 8.6	39.9 ± 10.7
XL all-comm	51.5 ± 2.3	37.5 ± 25.0	36.9 ± 18.1	26.8 ± 24.2	50.7 ± 9.4	40.7 ± 10.4
XL en-comm	49.9 ± 1.0	37.5 ± 25.0	38.3 ± 16.0	27.2 ± 24.3	50.3 ± 9.8	40.6 ± 9.7
XXL all-res	51.5 ± 2.8	38.2 ± 24.2	40.9 ± 18.5	45.3 ± 19.0	56.4 ± 10.4	46.5 ± 7.5
XXL en-res	53.4 ± 5.4	39.0 ± 24.8	40.1 ± 20.5	45.4 ± 20.6	59.0 ± 9.9	47.4 ± 8.7
XXL all-comm	50.6 ± 1.4	38.3 ± 24.3	45.2 ± 22.4	41.0 ± 20.2	58.9 ± 8.7	46.8 ± 8.2
XXL en-comm	52.5 ± 4.1	33.3 ± 27.0	43.9 ± 24.8	47.2 ± 17.8	59.2 ± 16.2	47.2 ± 9.7

Table 10: Flan-T5 models finetuned on two different instruction style configurations.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Small 1-english	28.3 ± 22.1	37.5 ± 25.0	35.3 ± 20.2	21.8 ± 26.5	44.8 ± 17.9	33.6 ± 8.8
Small 10-english	25.0 ± 22.0	38.1 ± 25.4	33.1 ± 24.4	20.6 ± 26.4	40.7 ± 19.5	31.5 ± 8.5
Base 1-english	51.1 ± 6.2	39.0 ± 26.0	36.2 ± 21.6	43.6 ± 21.2	57.6 ± 14.7	45.5 ± 8.8
Base 10-english	50.3 ± 2.4	38.8 ± 25.9	34.0 ± 22.4	43.0 ± 21.1	54.1 ± 13.0	44.1 ± 8.2
XL 1-english	60.6 ± 11.1	42.5 ± 28.8	52.1 ± 24.4	55.0 ± 18.7	81.3 ± 11.1	58.3 ± 14.5
XL 10-english	65.7 ± 15.2	45.1 ± 30.3	49.0 ± 23.5	56.8 ± 18.8	79.0 ± 11.4	59.1 ± 13.6
XXL 1-english	63.0 ± 13.1	43.9 ± 29.7	59.0 ± 30.5	58.1 ± 20.2	80.7 ± 9.9	60.9 ± 13.2
XXL 10-english	55.2 ± 23.7	46.3 ± 31.6	56.1 ± 29.1	57.7 ± 19.8	84.6 ± 9.6	60.0 ± 14.4

Table 11: mT5 models finetuned on three different instruction style configurations.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Small 1-english	30.2 ± 20.4	39.4 ± 25.1	35.0 ± 24.3	18.3 ± 24.2	37.8 ± 24.4	32.2 ± 8.5
Small 10-english	50.8 ± 3.1	38.4 ± 25.7	33.8 ± 23.6	17.9 ± 23.9	36.0 ± 23.0	35.4 ± 11.8
Small 10-multi	46.5 ± 13.4	39.4 ± 25.1	33.4 ± 23.5	18.2 ± 24.2	36.9 ± 23.5	34.9 ± 10.5
Base 1-english	53.4 ± 5.7	37.5 ± 23.8	34.7 ± 23.7	26.3 ± 23.7	44.3 ± 20.0	39.2 ± 10.2
Base 10-english	52.4 ± 5.1	37.3 ± 23.6	38.0 ± 17.9	21.8 ± 23.0	41.5 ± 20.5	38.2 ± 11.0
Base 10-multi	51.3 ± 3.2	38.0 ± 24.1	34.4 ± 22.7	29.6 ± 21.2	41.7 ± 18.1	39.0 ± 8.2
XL 1-english	51.7 ± 3.4	38.0 ± 24.1	36.9 ± 18.1	36.3 ± 21.7	50.9 ± 8.9	42.7 ± 7.8
XL 10-english	43.6 ± 16.5	38.0 ± 24.1	36.9 ± 18.1	30.9 ± 20.0	45.6 ± 13.8	39.0 ± 5.8
XL 10-multi	51.2 ± 3.3	38.0 ± 24.1	36.9 ± 18.1	31.1 ± 25.4	54.8 ± 12.9	42.4 ± 10.1



(a) Dataset Overlap

(b) Task Overlap

Figure 14: Difference to the baseline for the XL model across categories on LegalBench with dataset and task overlap held out respectively.

Table 12: Flan-T5 Small models with different domain adaptation strategies (amount of IFT data during continued pretraining). 1-IFT-to-X-PRE means that for every X pretraining examples we mix in one instruction example. ONLY-PRE means we did not mix in any instruction examples.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical	LegalBench
Baseline	0.3 ± 0.7	30.4 ± 20.3	39.8 ± 20.8	28.2 ± 21.6	27.7 ± 21.9	25.3 ± 13.2
IFT	25.0 ± 22.0	38.1 ± 25.4	43.0 ± 17.1	36.1 ± 26.5	32.6 ± 24.2	34.9 ± 6.0
1-IFT-to-200-PRE+IFT 10K	50.6 ± 4.2	38.2 ± 25.6	44.3 ± 15.6	33.8 ± 23.3	33.7 ± 23.8	40.1 ± 6.5
1-IFT-to-200-PRE+IFT 20K	50.8 ± 2.2	37.9 ± 25.3	44.4 ± 15.7	35.5 ± 25.1	31.9 ± 24.0	40.1 ± 6.7
1-IFT-to-200-PRE+IFT 30K	42.2 ± 16.2	37.3 ± 24.9	39.8 ± 19.4	34.3 ± 23.7	32.4 ± 23.5	37.2 ± 3.6
1-IFT-to-200-PRE+IFT 40K	45.8 ± 10.8	37.7 ± 25.2	39.7 ± 20.8	35.1 ± 24.4	33.4 ± 24.0	38.3 ± 4.3
1-IFT-to-200-PRE+IFT 50K	47.0 ± 8.8	37.4 ± 24.9	38.9 ± 20.7	35.6 ± 24.6	34.1 ± 21.0	38.6 ± 4.5
1-IFT-to-200-PRE+IFT 60K	50.0 ± 0.4	37.1 ± 24.7	39.3 ± 18.7	34.7 ± 23.3	33.8 ± 21.7	39.0 ± 5.8
1-IFT-to-200-PRE+IFT 70K	41.4 ± 16.9	38.4 ± 25.6	38.8 ± 21.1	34.0 ± 22.7	33.8 ± 22.9	37.3 ± 2.9
1-IFT-to-200-PRE+IFT 80K	51.8 ± 3.8	38.2 ± 25.5	38.5 ± 20.9	36.2 ± 22.6	33.4 ± 21.5	39.6 ± 6.4
1-IFT-to-200-PRE+IFT 90K	42.4 ± 16.7	37.9 ± 25.3	39.7 ± 20.3	35.8 ± 23.5	34.1 ± 22.2	38.0 ± 2.9
1-IFT-to-1000-PRE+IFT 10K	42.3 ± 16.1	38.1 ± 25.4	43.9 ± 15.0	33.6 ± 23.8	32.7 ± 24.5	38.1 ± 4.5
1-IFT-to-1000-PRE+IFT 20K	41.7 ± 20.5	37.0 ± 24.7	42.9 ± 16.6	33.1 ± 23.4	33.0 ± 24.6	37.5 ± 4.2
1-IFT-to-1000-PRE+IFT 30K	49.9 ± 0.4	37.8 ± 25.3	40.3 ± 17.7	34.3 ± 24.2	32.4 ± 23.5	38.9 ± 6.1
1-IFT-to-1000-PRE+IFT 40K	51.4 ± 2.7	37.8 ± 25.2	38.9 ± 20.6	34.7 ± 24.4	33.0 ± 22.5	39.2 ± 6.5
1-IFT-to-1000-PRE+IFT 50K	51.6 ± 2.7	37.7 ± 25.2	39.8 ± 18.4	33.7 ± 23.3	33.8 ± 22.4	39.3 ± 6.6
1-IFT-to-1000-PRE+IFT 60K	50.0 ± 0.6	37.5 ± 25.0	40.5 ± 20.2	34.4 ± 23.5	33.2 ± 22.4	39.1 ± 6.0
1-IFT-to-1000-PRE+IFT 70K	50.3 ± 1.4	37.3 ± 24.9	43.1 ± 17.1	34.6 ± 24.6	33.1 ± 22.4	39.7 ± 6.3
1-IFT-to-1000-PRE+IFT 80K	50.6 ± 1.5	37.7 ± 25.2	43.0 ± 17.4	34.0 ± 23.1	32.9 ± 23.0	39.6 ± 6.5
1-IFT-to-1000-PRE+IFT 90K	51.6 ± 2.6	37.0 ± 24.7	40.2 ± 19.2	34.4 ± 24.8	32.9 ± 21.4	39.2 ± 6.7
1-IFT-to-10000-PRE+IFT 10K	46.0 ± 12.1	38.0 ± 25.4	44.4 ± 15.5	33.5 ± 23.3	33.8 ± 24.3	39.1 ± 5.2
1-IFT-to-10000-PRE+IFT 20K	50.5 ± 1.4	37.9 ± 25.3	44.3 ± 15.4	34.9 ± 25.2	32.1 ± 24.0	39.9 ± 6.7
1-IFT-to-10000-PRE+IFT 30K	51.3 ± 4.0	38.2 ± 25.5	40.5 ± 18.1	33.6 ± 23.3	34.7 ± 26.5	39.7 ± 6.3
1-IFT-to-10000-PRE+IFT 40K	52.3 ± 4.4	38.9 ± 26.1	38.8 ± 19.8	33.2 ± 23.0	33.6 ± 25.3	39.4 ± 6.9
1-IFT-to-10000-PRE+IFT 50K	47.3 ± 12.3	37.6 ± 25.1	41.5 ± 17.2	35.1 ± 24.4	32.8 ± 22.2	38.8 ± 5.1
1-IFT-to-10000-PRE+IFT 60K	49.4 ± 2.7	38.1 ± 25.5	39.0 ± 20.6	35.3 ± 24.3	32.2 ± 23.2	38.8 ± 5.8
1-IFT-to-10000-PRE+IFT 70K	49.2 ± 13.9	37.7 ± 25.2	42.1 ± 16.2	33.2 ± 23.1	33.8 ± 24.3	39.2 ± 5.9
1-IFT-to-10000-PRE+IFT 80K	51.4 ± 7.0	37.5 ± 25.0	42.5 ± 16.0	33.5 ± 22.4	32.7 ± 22.4	39.5 ± 6.9
1-IFT-to-10000-PRE+IFT 90K	44.1 ± 20.2	37.5 ± 25.0	43.0 ± 16.4	33.6 ± 22.3	33.0 ± 21.9	38.2 ± 4.6
ONLY-PRE+IFT 10K	51.1 ± 3.1	37.9 ± 25.3	44.9 ± 16.9	33.8 ± 23.6	34.6 ± 24.7	40.5 ± 6.6
ONLY-PRE+IFT 20K	51.4 ± 4.4	38.1 ± 25.5	43.9 ± 14.0	34.1 ± 25.1	33.2 ± 25.3	40.2 ± 6.8
ONLY-PRE+IFT 30K	43.0 ± 17.8	37.9 ± 25.4	42.2 ± 16.2	35.1 ± 25.6	32.4 ± 23.6	38.1 ± 4.1
ONLY-PRE+IFT 40K	47.1 ± 12.5	38.4 ± 25.6	42.5 ± 16.6	34.9 ± 25.0	32.9 ± 24.5	39.2 ± 5.1
ONLY-PRE+IFT 50K	42.0 ± 19.2	37.8 ± 25.2	42.3 ± 17.4	34.8 ± 25.1	32.4 ± 23.3	37.8 ± 3.9
ONLY-PRE+IFT 60K	50.6 ± 2.1	37.9 ± 25.3	43.0 ± 16.0	35.6 ± 25.0	32.6 ± 22.9	39.9 ± 6.3
ONLY-PRE+IFT 70K	48.6 ± 7.0	38.1 ± 25.4	42.6 ± 17.0	34.8 ± 24.3	32.6 ± 24.0	39.4 ± 5.7
ONLY-PRE+IFT 80K	51.2 ± 3.4	37.5 ± 25.0	43.7 ± 17.2	33.2 ± 23.1	34.0 ± 25.7	39.9 ± 6.7
ONLY-PRE+IFT 90K	51.5 ± 3.7	37.5 ± 25.0	40.7 ± 17.5	34.7 ± 21.8	33.7 ± 24.4	39.6 ± 6.4

Table 13: Flan-T5 Base models with different domain adaptation strategies (amount of IFT data during continued pretraining). 1-IFT-to-X-PRE means that for every X pretraining examples we mix in one instruction example. ONLY-PRE means we did not mix in any instruction examples.

LLM	**Issue**	**Rule**	**Conclusion**	**Interpretation**	**Rhetorical**	**LegalBench**
Baseline	44.7 ± 12.4	18.0 ± 23.6	20.9 ± 24.8	28.9 ± 21.2	37.0 ± 21.3	29.9 ± 9.9
IFT	50.3 ± 2.4	38.8 ± 25.9	40.5 ± 15.7	49.5 ± 19.1	45.2 ± 22.0	44.9 ± 4.6
1-IFT-to-200-PRE+IFT 10K	50.5 ± 3.2	37.3 ± 24.9	40.7 ± 16.6	47.7 ± 17.7	49.7 ± 20.8	45.2 ± 5.2
1-IFT-to-200-PRE+IFT 20K	50.4 ± 2.2	37.8 ± 25.2	40.9 ± 14.2	48.4 ± 15.9	46.2 ± 24.7	44.7 ± 4.7
1-IFT-to-200-PRE+IFT 30K	49.9 ± 2.6	37.7 ± 25.2	41.2 ± 14.1	45.3 ± 16.1	48.4 ± 20.0	44.5 ± 4.5
1-IFT-to-200-PRE+IFT 40K	49.4 ± 4.3	37.8 ± 25.2	40.4 ± 15.5	47.8 ± 17.2	49.0 ± 20.9	44.9 ± 4.8
1-IFT-to-200-PRE+IFT 50K	51.2 ± 3.9	37.7 ± 25.2	41.2 ± 12.7	45.0 ± 16.0	49.1 ± 20.1	44.8 ± 4.9
1-IFT-to-200-PRE+IFT 60K	50.1 ± 0.9	37.6 ± 25.1	45.1 ± 13.0	44.2 ± 16.0	45.2 ± 18.9	44.4 ± 4.0
1-IFT-to-200-PRE+IFT 70K	51.1 ± 2.7	37.6 ± 25.0	43.4 ± 13.6	45.1 ± 15.4	46.5 ± 21.0	44.7 ± 4.4
1-IFT-to-200-PRE+IFT 80K	50.4 ± 2.3	37.7 ± 25.2	42.2 ± 15.9	45.2 ± 15.7	44.8 ± 22.7	44.1 ± 4.1
1-IFT-to-200-PRE+IFT 90K	51.4 ± 3.6	37.7 ± 25.2	41.6 ± 14.5	42.9 ± 19.0	43.2 ± 21.6	43.4 ± 4.5
1-IFT-to-1000-PRE+IFT 10K	46.8 ± 4.8	38.5 ± 25.7	43.9 ± 13.7	47.6 ± 16.6	45.9 ± 18.0	44.5 ± 3.2
1-IFT-to-1000-PRE+IFT 20K	50.1 ± 2.0	37.8 ± 25.2	43.2 ± 15.0	46.7 ± 15.9	48.2 ± 24.9	45.2 ± 4.3
1-IFT-to-1000-PRE+IFT 30K	50.8 ± 3.3	38.9 ± 26.0	42.3 ± 15.9	49.9 ± 17.6	50.4 ± 21.4	46.5 ± 4.9
1-IFT-to-1000-PRE+IFT 40K	50.1 ± 0.7	38.4 ± 25.7	45.1 ± 12.0	46.6 ± 16.2	48.0 ± 21.4	45.7 ± 4.0
1-IFT-to-1000-PRE+IFT 50K	51.1 ± 3.0	37.7 ± 25.1	41.9 ± 13.8	48.0 ± 19.3	50.1 ± 20.5	45.8 ± 5.1
1-IFT-to-1000-PRE+IFT 60K	49.9 ± 2.3	37.7 ± 25.1	44.2 ± 15.7	46.1 ± 18.3	49.7 ± 22.1	45.5 ± 4.5
1-IFT-to-1000-PRE+IFT 70K	50.5 ± 1.5	38.5 ± 25.7	44.9 ± 16.8	47.9 ± 15.9	49.8 ± 19.2	46.3 ± 4.4
1-IFT-to-1000-PRE+IFT 80K	50.6 ± 2.5	37.9 ± 25.2	42.4 ± 16.6	48.8 ± 19.2	48.7 ± 22.8	45.7 ± 4.8
1-IFT-to-1000-PRE+IFT 90K	50.8 ± 4.2	37.8 ± 25.2	43.4 ± 15.7	45.9 ± 16.9	47.8 ± 22.0	45.1 ± 4.4
1-IFT-to-10000-PRE+IFT 10K	48.8 ± 4.1	38.1 ± 25.4	43.6 ± 13.4	47.4 ± 16.4	47.7 ± 19.6	45.1 ± 3.9
1-IFT-to-10000-PRE+IFT 20K	50.0 ± 2.9	37.7 ± 25.1	41.5 ± 13.6	47.2 ± 18.4	52.0 ± 20.8	45.7 ± 5.3
1-IFT-to-10000-PRE+IFT 30K	50.5 ± 4.6	38.4 ± 25.6	44.3 ± 14.6	48.4 ± 17.3	51.5 ± 20.7	46.6 ± 4.8
1-IFT-to-10000-PRE+IFT 40K	50.2 ± 2.9	37.7 ± 25.1	42.4 ± 16.4	45.6 ± 16.8	49.2 ± 20.7	45.0 ± 4.6
1-IFT-to-10000-PRE+IFT 50K	50.3 ± 2.0	37.4 ± 24.9	41.8 ± 16.2	45.8 ± 17.7	49.3 ± 21.7	44.9 ± 4.8
1-IFT-to-10000-PRE+IFT 60K	49.6 ± 4.5	37.6 ± 25.1	43.7 ± 17.3	43.1 ± 19.3	48.4 ± 22.0	44.5 ± 4.3
1-IFT-to-10000-PRE+IFT 70K	49.6 ± 2.9	37.7 ± 25.1	46.4 ± 16.0	46.9 ± 18.7	50.5 ± 22.2	46.2 ± 4.5
1-IFT-to-10000-PRE+IFT 80K	49.7 ± 3.0	37.7 ± 25.2	45.1 ± 12.2	41.1 ± 18.4	47.7 ± 23.7	44.2 ± 4.3
1-IFT-to-10000-PRE+IFT 90K	50.0 ± 1.8	37.2 ± 24.8	40.6 ± 14.5	41.8 ± 20.0	45.3 ± 22.3	43.0 ± 4.4
ONLY-PRE+IFT 10K	50.7 ± 2.7	37.2 ± 24.8	42.0 ± 16.3	48.0 ± 18.6	47.6 ± 20.8	45.1 ± 4.9
ONLY-PRE+IFT 20K	50.1 ± 2.6	38.2 ± 25.5	41.1 ± 13.7	45.0 ± 19.7	46.7 ± 25.7	44.2 ± 4.2
ONLY-PRE+IFT 30K	50.7 ± 3.6	38.0 ± 25.3	43.3 ± 15.3	44.6 ± 19.0	48.3 ± 21.6	45.0 ± 4.4
ONLY-PRE+IFT 40K	50.4 ± 3.8	38.4 ± 25.6	41.9 ± 14.5	47.4 ± 17.4	46.8 ± 21.4	45.0 ± 4.3
ONLY-PRE+IFT 50K	50.6 ± 2.5	37.5 ± 25.0	41.1 ± 12.8	44.5 ± 18.6	48.2 ± 21.6	44.4 ± 4.7
ONLY-PRE+IFT 60K	49.6 ± 3.4	37.6 ± 25.1	40.4 ± 15.5	47.2 ± 16.7	46.3 ± 21.0	44.2 ± 4.5
ONLY-PRE+IFT 70K	50.6 ± 1.9	38.4 ± 25.6	41.7 ± 13.2	46.1 ± 18.7	45.5 ± 21.9	44.4 ± 4.2
ONLY-PRE+IFT 80K	51.0 ± 3.1	39.2 ± 26.3	42.2 ± 15.7	46.8 ± 18.0	45.3 ± 21.9	44.9 ± 4.0
ONLY-PRE+IFT 90K	50.5 ± 3.8	37.4 ± 25.0	44.3 ± 14.7	43.2 ± 18.1	44.4 ± 22.5	44.0 ± 4.1

Table 14: Flan-T5 XL models with different domain adaptation strategies (amount of IFT data during continued pretraining). 1-IFT-to-X-PRE means that for every X pretraining examples we mix in one instruction example. ONLY-PRE means we did not mix in any instruction examples.

LLM	**Issue**	**Rule**	**Conclusion**	**Interpretation**	**Rhetorical**	**LegalBench**
Baseline	53.5 ± 6.0	32.1 ± 24.6	46.8 ± 15.6	58.7 ± 21.3	59.6 ± 25.6	50.1 ± 10.1
IFT	65.7 ± 15.2	45.1 ± 30.3	49.5 ± 14.2	61.7 ± 17.1	68.6 ± 24.1	58.1 ± 9.2
1-IFT-to-200-PRE+IFT 10K	56.7 ± 6.9	41.8 ± 28.1	55.2 ± 16.9	62.1 ± 18.6	66.8 ± 23.7	56.5 ± 8.4
1-IFT-to-200-PRE+IFT 20K	63.4 ± 13.8	44.2 ± 29.8	52.5 ± 17.4	58.7 ± 16.8	67.0 ± 23.2	57.2 ± 8.1
1-IFT-to-200-PRE+IFT 30K	58.7 ± 10.3	43.6 ± 29.3	56.3 ± 18.2	60.2 ± 18.4	67.9 ± 24.5	57.3 ± 7.9
1-IFT-to-200-PRE+IFT 40K	58.4 ± 9.7	42.3 ± 28.2	54.3 ± 15.2	61.2 ± 18.8	67.5 ± 23.6	56.7 ± 8.4
1-IFT-to-200-PRE+IFT 50K	61.4 ± 13.3	42.2 ± 28.3	51.8 ± 16.3	59.4 ± 17.9	67.3 ± 23.6	56.4 ± 8.7
1-IFT-to-200-PRE+IFT 60K	57.5 ± 8.7	43.6 ± 29.2	53.5 ± 15.8	60.3 ± 17.5	68.2 ± 23.5	56.6 ± 8.1
1-IFT-to-200-PRE+IFT 70K	58.3 ± 10.2	43.1 ± 28.8	54.3 ± 17.9	58.8 ± 18.2	67.6 ± 22.6	56.4 ± 8.0
1-IFT-to-200-PRE+IFT 80K	58.9 ± 11.0	44.9 ± 30.0	51.1 ± 13.2	59.8 ± 17.3	68.5 ± 23.3	56.6 ± 8.1
1-IFT-to-200-PRE+IFT 90K	55.2 ± 6.9	44.4 ± 30.1	51.7 ± 15.6	57.9 ± 17.0	67.7 ± 24.3	55.4 ± 7.6
1-IFT-to-1000-PRE+IFT 10K	61.3 ± 11.8	41.8 ± 28.0	53.4 ± 16.1	60.9 ± 18.8	67.0 ± 23.1	56.9 ± 8.7
1-IFT-to-1000-PRE+IFT 20K	63.3 ± 13.7	44.3 ± 29.6	52.2 ± 17.4	60.7 ± 17.5	67.3 ± 24.6	57.6 ± 8.3
1-IFT-to-1000-PRE+IFT 30K	58.3 ± 9.8	43.4 ± 29.2	54.4 ± 17.1	61.3 ± 20.4	70.2 ± 25.4	57.5 ± 8.8
1-IFT-to-1000-PRE+IFT 40K	62.5 ± 13.2	45.6 ± 30.6	51.3 ± 17.5	60.1 ± 18.9	68.0 ± 25.6	57.5 ± 8.0
1-IFT-to-1000-PRE+IFT 50K	56.8 ± 7.5	44.7 ± 30.2	51.5 ± 14.5	58.9 ± 16.9	69.7 ± 24.9	56.3 ± 8.3
1-IFT-to-1000-PRE+IFT 60K	54.4 ± 5.3	42.2 ± 28.2	52.7 ± 16.3	59.9 ± 17.8	67.1 ± 23.5	55.2 ± 8.2
1-IFT-to-1000-PRE+IFT 70K	59.7 ± 10.8	44.1 ± 29.5	54.5 ± 17.3	59.4 ± 17.6	67.4 ± 23.4	57.0 ± 7.7
1-IFT-to-1000-PRE+IFT 80K	59.8 ± 11.2	41.6 ± 27.9	52.8 ± 17.2	63.5 ± 19.8	67.3 ± 24.5	57.0 ± 9.0
1-IFT-to-1000-PRE+IFT 90K	60.3 ± 10.6	44.3 ± 29.7	50.5 ± 15.4	57.3 ± 15.9	67.3 ± 23.1	55.9 ± 8.0
1-IFT-to-10000-PRE+IFT 10K	60.0 ± 10.2	42.3 ± 28.4	52.7 ± 16.0	61.6 ± 18.3	68.0 ± 22.8	56.9 ± 8.8
1-IFT-to-10000-PRE+IFT 20K	59.5 ± 11.0	42.6 ± 28.5	52.5 ± 15.7	61.6 ± 18.0	68.1 ± 25.0	56.9 ± 8.7
1-IFT-to-10000-PRE+IFT 30K	62.2 ± 12.2	42.3 ± 28.5	53.6 ± 16.7	62.5 ± 20.1	69.2 ± 25.2	57.9 ± 9.3
1-IFT-to-10000-PRE+IFT 40K	59.7 ± 10.1	43.6 ± 29.2	53.1 ± 15.9	62.6 ± 18.9	67.6 ± 23.1	57.3 ± 8.3
1-IFT-to-10000-PRE+IFT 50K	58.8 ± 8.9	42.9 ± 29.1	52.5 ± 16.9	61.1 ± 17.9	64.6 ± 25.0	56.0 ± 7.6
1-IFT-to-10000-PRE+IFT 60K	55.3 ± 5.6	42.1 ± 28.3	52.1 ± 16.6	59.1 ± 19.0	66.4 ± 23.1	55.0 ± 8.0
1-IFT-to-10000-PRE+IFT 70K	60.3 ± 10.0	43.6 ± 29.5	51.8 ± 16.8	61.2 ± 18.5	69.0 ± 24.7	57.2 ± 8.7
1-IFT-to-10000-PRE+IFT 80K	64.7 ± 13.9	44.4 ± 29.9	50.8 ± 16.9	58.4 ± 17.1	70.4 ± 25.8	57.8 ± 9.3
1-IFT-to-10000-PRE+IFT 90K	63.3 ± 13.3	44.8 ± 30.2	51.9 ± 16.3	58.7 ± 16.6	68.2 ± 25.1	57.4 ± 8.3
ONLY-PRE+IFT 10K	62.8 ± 13.6	44.3 ± 29.8	52.0 ± 16.7	58.9 ± 16.2	68.2 ± 23.9	57.2 ± 8.3
ONLY-PRE+IFT 20K	64.0 ± 13.9	42.6 ± 28.7	52.8 ± 15.6	62.0 ± 18.0	68.7 ± 25.0	58.0 ± 9.3
ONLY-PRE+IFT 30K	52.9 ± 15.5	42.0 ± 28.3	51.5 ± 16.0	62.0 ± 18.7	67.3 ± 24.9	55.1 ± 8.8
ONLY-PRE+IFT 40K	60.4 ± 12.2	43.1 ± 29.1	52.4 ± 16.9	60.6 ± 17.5	68.9 ± 23.4	57.1 ± 8.7
ONLY-PRE+IFT 50K	57.4 ± 8.5	42.6 ± 28.8	51.6 ± 15.3	61.2 ± 18.1	70.0 ± 23.8	56.5 ± 9.2
ONLY-PRE+IFT 60K	56.7 ± 7.6	42.5 ± 28.4	52.0 ± 16.3	61.2 ± 17.9	68.8 ± 23.8	56.2 ± 8.8
ONLY-PRE+IFT 70K	57.2 ± 8.5	42.1 ± 28.4	51.5 ± 17.0	60.8 ± 18.1	70.2 ± 24.8	56.3 ± 9.4
ONLY-PRE+IFT 80K	60.3 ± 11.1	42.4 ± 28.4	54.6 ± 16.4	65.1 ± 20.9	69.2 ± 24.8	58.3 ± 9.3
ONLY-PRE+IFT 90K	60.3 ± 12.0	44.4 ± 29.8	52.3 ± 17.1	59.8 ± 17.8	67.8 ± 24.4	56.9 ± 7.9

Table 15: Flan-T5 XXL models with different domain adaptation strategies (amount of IFT data during continued pretraining). 1-IFT-to-X-PRE means that for every X pretraining examples we mix in one instruction example. ONLY-PRE means we did not mix in any instruction examples.

LLM	**Issue**	**Rule**	**Conclusion**	**Interpretation**	**Rhetorical**	**LegalBench**
Baseline	36.1 ± 21.5	18.8 ± 24.6	25.2 ± 26.0	35.1 ± 22.2	41.1 ± 18.4	31.3 ± 8.1
IFT	55.2 ± 23.7	46.3 ± 31.6	56.2 ± 18.3	66.3 ± 19.7	73.8 ± 24.4	59.6 ± 9.5
1-IFT-to-200-PRE+IFT 10K	53.4 ± 16.2	47.9 ± 32.1	58.1 ± 19.5	63.8 ± 17.6	74.2 ± 27.1	59.5 ± 9.0
1-IFT-to-200-PRE+IFT 20K	53.6 ± 3.7	48.9 ± 32.9	58.8 ± 18.7	65.3 ± 17.5	72.0 ± 25.5	59.7 ± 8.2
1-IFT-to-200-PRE+IFT 30K	56.5 ± 18.3	48.9 ± 31.5	60.5 ± 19.9	65.2 ± 18.3	69.5 ± 24.2	60.1 ± 7.1
1-IFT-to-200-PRE+IFT 40K	58.3 ± 20.2	47.3 ± 30.8	57.9 ± 19.1	65.6 ± 18.2	71.3 ± 24.1	60.1 ± 8.1
1-IFT-to-200-PRE+IFT 50K	60.3 ± 12.6	48.4 ± 31.4	63.2 ± 20.2	67.9 ± 18.9	71.4 ± 26.1	62.2 ± 7.9
1-IFT-to-200-PRE+IFT 60K	58.6 ± 20.5	48.5 ± 31.5	60.9 ± 20.7	67.5 ± 19.9	71.0 ± 24.7	61.3 ± 7.8
1-IFT-to-200-PRE+IFT 70K	58.6 ± 10.5	48.5 ± 31.4	60.6 ± 20.4	65.3 ± 18.4	69.3 ± 23.4	60.5 ± 7.0
1-IFT-to-200-PRE+IFT 80K	53.7 ± 16.4	47.8 ± 30.8	58.8 ± 18.2	63.7 ± 17.7	71.3 ± 25.7	59.1 ± 8.1
1-IFT-to-200-PRE+IFT 90K	52.0 ± 14.5	48.8 ± 31.7	59.4 ± 19.6	64.4 ± 17.9	72.3 ± 25.1	59.4 ± 8.5
1-IFT-to-1000-PRE+IFT 10K	41.1 ± 24.2	45.9 ± 30.3	58.2 ± 18.4	65.5 ± 20.2	68.8 ± 25.2	55.9 ± 10.8
1-IFT-to-1000-PRE+IFT 20K	47.7 ± 24.8	48.0 ± 31.1	60.3 ± 20.3	67.2 ± 19.7	70.3 ± 23.8	58.7 ± 9.4
1-IFT-to-1000-PRE+IFT 30K	40.3 ± 28.4	45.5 ± 29.6	62.3 ± 21.1	67.8 ± 21.1	69.3 ± 22.6	57.0 ± 11.9
1-IFT-to-1000-PRE+IFT 40K	44.2 ± 27.4	46.7 ± 29.9	61.9 ± 21.9	68.6 ± 20.7	71.2 ± 24.9	58.5 ± 11.1
1-IFT-to-1000-PRE+IFT 50K	49.7 ± 25.2	49.1 ± 33.1	55.5 ± 19.2	68.2 ± 19.8	71.4 ± 24.3	58.8 ± 9.3
1-IFT-to-1000-PRE+IFT 60K	44.9 ± 22.0	47.6 ± 30.7	57.9 ± 19.4	69.7 ± 21.1	72.1 ± 26.0	58.5 ± 11.1
1-IFT-to-1000-PRE+IFT 70K	40.6 ± 25.0	48.1 ± 31.2	60.5 ± 20.0	68.2 ± 20.5	72.5 ± 24.4	58.0 ± 12.0
1-IFT-to-1000-PRE+IFT 80K	53.8 ± 23.7	47.9 ± 32.4	53.5 ± 17.5	67.1 ± 19.3	71.8 ± 25.9	58.8 ± 9.1
1-IFT-to-1000-PRE+IFT 90K	47.6 ± 23.5	47.1 ± 30.5	60.1 ± 18.9	65.1 ± 24.3	70.3 ± 23.5	58.0 ± 9.3
1-IFT-to-10000-PRE+IFT 10K	49.8 ± 13.6	46.6 ± 30.0	59.0 ± 16.6	64.6 ± 19.3	72.6 ± 24.7	58.5 ± 9.5
1-IFT-to-10000-PRE+IFT 20K	45.2 ± 27.4	46.3 ± 31.2	58.8 ± 20.1	68.1 ± 19.0	71.7 ± 24.1	58.0 ± 10.9
1-IFT-to-10000-PRE+IFT 30K	46.8 ± 24.6	46.0 ± 29.6	62.6 ± 18.4	66.1 ± 18.1	72.1 ± 25.3	58.7 ± 10.5
1-IFT-to-10000-PRE+IFT 40K	56.8 ± 24.5	46.9 ± 30.4	59.1 ± 19.3	68.3 ± 21.1	72.2 ± 26.2	60.7 ± 8.9
1-IFT-to-10000-PRE+IFT 50K	54.5 ± 28.7	43.1 ± 28.1	62.2 ± 19.8	64.2 ± 19.1	70.2 ± 24.3	58.8 ± 9.3
1-IFT-to-10000-PRE+IFT 60K	52.0 ± 16.0	42.0 ± 28.7	60.3 ± 17.4	65.7 ± 19.6	71.3 ± 24.7	58.2 ± 10.3
1-IFT-to-10000-PRE+IFT 70K	52.2 ± 14.7	47.4 ± 30.8	59.2 ± 18.3	66.6 ± 18.5	70.0 ± 24.1	59.1 ± 8.5
1-IFT-to-10000-PRE+IFT 80K	56.5 ± 18.5	44.9 ± 28.9	59.7 ± 17.2	65.3 ± 17.7	72.3 ± 25.6	59.7 ± 9.1
1-IFT-to-10000-PRE+IFT 90K	45.0 ± 17.4	41.5 ± 27.3	56.3 ± 16.3	66.3 ± 18.5	72.1 ± 25.7	56.2 ± 11.8
ONLY-PRE+IFT 10K	49.2 ± 24.4	47.1 ± 30.4	62.0 ± 20.3	66.9 ± 20.4	71.7 ± 25.1	59.4 ± 9.7
ONLY-PRE+IFT 20K	35.6 ± 24.0	46.2 ± 30.0	56.3 ± 17.9	62.3 ± 18.4	68.6 ± 24.2	53.8 ± 11.7
ONLY-PRE+IFT 30K	46.3 ± 28.4	45.7 ± 29.3	56.1 ± 18.5	67.7 ± 19.9	72.1 ± 25.6	57.6 ± 10.8
ONLY-PRE+IFT 40K	48.8 ± 30.3	45.7 ± 29.5	56.6 ± 18.0	68.1 ± 20.0	71.6 ± 26.3	58.1 ± 10.2
ONLY-PRE+IFT 50K	47.5 ± 24.9	47.1 ± 30.2	53.5 ± 16.2	67.1 ± 19.5	71.8 ± 25.4	57.4 ± 10.2
ONLY-PRE+IFT 60K	33.2 ± 23.3	47.8 ± 30.7	55.0 ± 17.9	63.1 ± 19.7	69.3 ± 25.0	53.7 ± 12.6
ONLY-PRE+IFT 70K	42.7 ± 25.9	47.2 ± 30.5	55.9 ± 19.4	60.7 ± 17.5	68.0 ± 23.8	54.9 ± 9.1
ONLY-PRE+IFT 80K	43.7 ± 25.8	46.3 ± 29.9	55.8 ± 17.1	64.8 ± 18.7	71.8 ± 25.9	56.5 ± 10.7
ONLY-PRE+IFT 90K	55.3 ± 16.9	45.2 ± 28.9	60.0 ± 17.0	64.9 ± 20.0	69.0 ± 24.3	58.9 ± 8.2