LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text

Dor Bernsohn[⋆] **Gil Semo**[⋆] **Yaron Vazana**[⋆] **Gila Hayat**[⋆] **Ben Hagag**[⋆]
Joel Niklaus[†] **Rohit Saha**[‡] **Kyryl Truskovskyi**[‡]

[⋆]Darrow AI Ltd., Tel Aviv, Israel {firstname.lastname}@darrow.ai
[†]Niklaus.ai, Bern, Switzerland joel@niklaus.ai
[‡]Georgian.io, Toronto, Canada {firstname}@georgian.io

Abstract

In this study, we focus on two main tasks, the first for detecting legal violations within unstructured textual data, and the second for associating these violations with potentially affected individuals. We constructed two datasets using Large Language Models (LLMs) which were subsequently validated by domain expert annotators. Both tasks were designed specifically for the context of class-action cases. The experimental design incorporated fine-tuning models from the BERT family and open-source LLMs, and conducting few-shot experiments using closed-source LLMs. Our results, with an F1-score of 62.69% (violation identification) and 81.02% (associating victims), show that our datasets and setups can be used for both tasks. Finally, we publicly release the datasets and the code used for the experiments in order to advance further research in the area of legal natural language processing (NLP).

1 Introduction

The widespread use of the internet has changed how information moves and connects in our society. Every day, the digital domain is flooded with a multitude of textual data, spanning from news articles and reviews to social media posts [1]. Within this sea of unstructured text, legal violations can often go unnoticed, concealed by the vast amount of surrounding information. These violations not only pose potential harm to individuals and entities but also challenge the very fabric of legal and ethical standards in the digital era. The significance of addressing these hidden violations cannot be overstated; as they have widespread implications for individual rights, societal norms, and the principles of justice. As a result, there is a pressing need to develop sophisticated methods to sift through the noise and identify these breaches.

[1]https://www.internetlivestats.com/total-number-of-websites

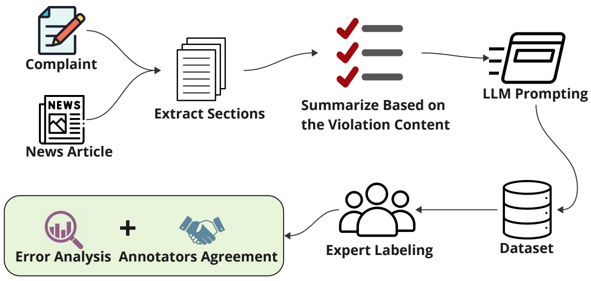

Figure 1: A visual representation of the data generation flow, illustrating the step-by-step process from raw input to the final synthesized dataset.

Legal violations often leave data trails. To detect these trails for pinpointing the violations, previous studies have often relied on specialized models tailored for specific domain applications (Silva et al., 2020; Yu et al., 2020). These models, while effective in their specific domains, lack the versatility needed to address the wide array of legal violations that can occur across different contexts.

Legal violation identification aims to automatically uncover legal violations from unstructured text sources and assign potential victims to these violations. We designed two setups, one for each task, the first for solving the legal violation identification task (a.k.a Identification Setup) using named entity recognition (NER), and the other for associating these violations with potentially affected individuals (a.k.a Resolution Setup) using natural language inference (NLI). Our dataset for the NER task is not limited to any specific domain, while the NLI dataset is focused on four common legal domains. Followed by recent research in the field of data generation (Leiker et al., 2023; Veselovsky et al., 2023; Hämäläinen et al., 2023), we chose to employ GPT-4 (OpenAI, 2023) for synthetic data generation due to his ability to produce a large, diverse, and high-quality dataset that closely mimics the syntactic complexity of legal language, offering a scalable and ethically sound alternative to manual

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 2129–2145
March 17-22, 2024

data crafting. We employed a thorough verification process to validate the data for both its realistic and complexity. Our approach involved automated data generation based on real-world event contexts in the English language, complemented by manual reviews conducted by seasoned legal annotators on the generated data.

Contributions

The contributions of this paper are three-fold:

- We introduce two dedicated datasets for legal violation identification, based on previous class action cases and legal news. These datasets, which include new legal entities, were generated using LLMs and validated by domain experts.
- We evaluate various language models, including BERT-based models and LLMs, across two different NLP tasks, offering valuable insights into their applicability and limitations in the context of legal NLP.
- We implement a two-setup approach employing both NER and NLI tasks, providing a methodology for legal violation detection and resolution.

Main Research Questions

We believe numerous violations exist in unstructured text. Our aim is to uncover these violations and link them to relevant prior class actions. This study focuses on the following key research questions:

RQ1: To what extent do our newly introduced datasets enhance the performance of language models in identifying legal violations within unstructured text and associate victims to them?

RQ2: How effectively do the language models adapt to new, unseen data for the purpose of identifying legal violations and correlating them with past resolved cases across different legal domains?

RQ3: What is the level of difference between machine-generated and human-generated text in the context of legal violation identification?

2 Related Work

Previous works in the field of legal violation identification mostly focused on domain-specific topics, encompassing areas such as compliance, data privacy, and industry-specific regulations. For instance, Amaral et al. (2023) evaluates data agreements for compliance with European privacy laws using NLP techniques. Silva et al. (2020) used NER to identify personal information in datasets, thereby uncovering instances of online data privacy breaches. Nyffenegger et al. (2023) used LLMs to attempt re-identification of anonymized persons from court decisions. Additionally, neural networks have been used to classify and annotate violation cases in specific industries like power supply (Yu et al., 2020). These studies, while valuable, have generally been limited to specific types of legal domains or particular sectors. Our work contributes to this existing body of research by introducing a dataset designed for broader applicability in identifying various types of legal violations.

Prior research has explored the use of Large Language Models (LLMs) for synthetic data generation (Rosenbaum et al., 2022a,b), beneficial in situations with scarce authentic data (Brown et al., 2020). In fact, training models on synthetic data led to improved outcomes in benchmarks like SQUAD1.1 (Puri et al., 2020). However, human-curated data often provides a richness that is hard to replicate (Møller et al., 2023; Ding et al., 2022). In this paper, we present a multi-step validation method to discern between real-world and machine-generated content, addressing the inherent limitations of relying solely on synthetic data.

Previous studies indicate that LLMs are capable of explaining legal terms present in legislative documents by drafting explanations of how previous courts explained the meaning of statutory terms (Savelka et al., 2023b). Moreover, the models demonstrated analytical depth in court decision analysis, rivaling seasoned law students (Savelka et al., 2023a). In this study, we created a dataset based on a previous lawsuits legislation background, rather than examining existing records.

While LLMs (Radford et al., 2019) have been employed to enhance datasets for event detection tasks (Veyseh et al., 2021), our methodology advances this by generating pairs of specific violations and their corresponding events, using data from previously settled lawsuits. Unlike Koreeda and Manning (2021), who concentrated on NLI in the context of legal contracts, our research introduces an NLI dataset based on class-action cases. Additionally, NER has been increasingly applied in the legal domain, including efforts to extract entities from Indian court judgments (Kalamkar et al., 2022) and other legal texts (Luz de Araujo et al., 2018; Angelidis et al., 2018; Leitner et al., 2019). Despite these advancements, existing research has largely focused on a standard set of entity types, such as parties (plaintiff and defendant), judges, court name and law/citation. Our work introduces a new set of entity types that have not been previously explored in legal NER research (Păiș et al., 2021; Luz de Araujo et al., 2018; Dozier et al., 2010; Leitner et al., 2020; Skylaki et al., 2020; Kalamkar et al., 2022), thereby expanding the scope and applicability of NER in legal contexts.

3 Curating Custom Legal Datasets: A Multi-stage Approach to NER and NLI Tasks

Existing datasets may not adequately address the diverse range of legal violations and contexts central to our study, which is not in specific areas. To overcome these challenges, we employed a systematic and carefully planned data generation process, consisting of three stages: prompting, labeling, and data validation. This approach aimed at creating two robust datasets for two NLP tasks in the legal domain. We chose to focus on two key tasks:

- NER (classifying tokens into predefined entities) for identifying violations. NER has been employed to define novel legal entities, enabling precise localization of pertinent information necessary for the extraction of legitimate legal violations, as detailed in Table 4 in Appendix C.
- NLI (classifying a hypothesis and a premise into entailed/contradict/neutral) for matching these violations with known, resolved class-action cases. NLI facilitates the correlation of multiple unstructured text associated with the same violation, thereby enabling the matching of extracted violations identified by the NER task with pre-existing legal complaints of class action cases.

This dual-setup approach was designed to mimic the process of legal violation detection and resolution, generating high-quality data that closely resembles real-world scenarios.

Based on recent research in prompt-based methods (Liu et al., 2023), our study employs prompts for a variety of reasons. LLMs have been shown to adapt to specialized tasks through techniques like instruction tuning (Wei et al., 2021), reinforcement learning from human feedback (Ouyang et al., 2022), and in-context learning (Brown et al., 2020) when prompted with natural language instructions. Prompts facilitate task-specific optimization, a quality emphasized by DialogPrompt (Gu et al., 2021), which aligns with our focus on NER and NLI in the legal domain by fine-tuning on the generated dataset. Additionally, the sensitivity of prompts in context, as demonstrated in Time-aware Prompts in Text Generation (Cao and Wang, 2022), is crucial for understanding specific legal contexts like resolved class-action cases. As a result, our methodology leverages a prompt-based approach, optimized for the legal domain, to generate high-quality data for NER and NLI tasks.

3.1 Interconnection Between NER and NLI

The process of identifying and resolving legal violations in unstructured text involves the collaborative use of NER and NLI. Initially, a NER model scans the text to detect 'VIOLATION' entities, and if a potential violation is tagged with a high-confidence score, it's considered for further analysis. Subsequently, the text is processed through an NLI model in a pair-wise fashion against a dataset of closed settlements. If the NLI model finds a logical entailment between the text and any of the settled cases, indicating a substantial similarity, the corresponding complaints are flagged as candidates for matching with the specific user's complaint, potentially qualifying them for inclusion in a settlement fund. This streamlined approach harnesses the strengths of both NER and NLI to efficiently identify and associate potential legal violations with relevant precedents.

3.2 NER Data Generation

NER can be framed as a token classification task, wherein, the objective is to classify each word in a sentence as an entity class. In our dataset, there are four such entities; *Law*, *Violation*, *Violated By*, and *Violated On*.

For the NER task, our foundational data source was class action complaints, as described in (Semo et al., 2022). A complaint, often referred to as a plaintiff's plea, is a formal legal document that initiates a lawsuit. It outlines the complaints of the plaintiff and specifies the relief sought from the court. From each of these complaints, we extracted relevant sections such as *allegations*, *counts*, and *legal arguments* that were pertinent to our study, ensuring relevance and precision. These sections encapsulate the main context of the alleged violations. They were subsequently summarized through the utilization of GPT-4 (OpenAI, 2023) to capture the core essence of the violation content, and were employed as the context in the subsequent prompts.

For a visual representation of our data generation process, refer to Figure 1.

Prompt

For the NER task, we devised two unique prompting strategies: explicit and implicit. The explicit method not only emphasizes the inclusion of multiple distinct entities but also underscores the specific order of their appearance, adding a layer of complexity and structure to the generated content (refer to figure 6 in the Appendix). This approach ensures that the content is not only diverse but also adheres to certain structural guidelines, which contain task descriptions, specific instructions, and few-shot examples. Conversely, the implicit strategy focuses solely on a singular entity, specifically the content that describes the violation, refer to figure 6 in the Appendix.

Furthermore, both strategies incorporate additional parameters such as the cause of action, industry, and context. The inclusion of these parameters refines the generated content, tailoring it to specific scenarios and ensuring its relevance to the desired domain. By employing the explicit approach, we capture the comprehensive nature of a scenario, whereas the implicit method provides a concise perspective on one specific aspect.

3.3 NLI Data Generation

NLI can be framed as a classification task, wherein, the objective is to compare a premise to a hypothesis, and predict one of the three classes: (1) *Entailment* - where the hypothesis is contained and can be supported by the premise, (2) *Contradiction* - when the hypothesis contradicts the premise, (3) *Neutral* - when the premise neither entails nor contradicts the hypothesis.

For the NLI task, our data source consisted articles taken from a legal news website. Each news article was first summarized, by prompting GPT-4 (OpenAI, 2023), to capture its legal grounds. By summarizing, we ensured that the data was concise yet comprehensive by keeping only the legal violation section and removing background parts. This summarized content served as the premise. Using this premise, the model was tasked to generate a hypothesis that mimicked real-world scenarios. The intention behind this design was to create diverse records that spanned various legal areas. Table 5 in Appendix C presents the NLI data distributions.

Prompt

In this setup, we aimed to create scenarios that mirror real-life accounts of potential violations. We generated texts that mimic common situations where individuals share concerns, like online reviews or social media posts. The goal was to produce narratives that implicitly describe the effects of a violation. We added variations in attributes such as the writers age and gender and the text format to capture a wide range of experiences.

4 Human Expert Annotations

Data validation holds particular importance in our study due to the synthetic nature of the dataset. To ensure that the dataset is both realistic and challenging, we have implemented several validation methods. In this structured process, summaries of complaint documents and tasks for the NER and NLI models were generated automatically. Legal experts then carefully examined these auto-generated summaries and tasks. Their primary role was to meticulously review each output, ensuring that the summaries accurately reflected the key points of the complaints and that the tasks were correctly aligned with the context provided by these summaries. Additionally, each record was subjected to examination by several annotators, which serves to reduce potential bias in the evaluation. These annotators were tasked with identifying and suggesting any missing entities, as well as in checking for hallucinations—instances where the generated content might stray from factual accuracy. To maintain a rigorous and unbiased validation, all annotators received identical instructions, and the data presented to them was systematically shuffled. Their detailed examination was crucial in pinpointing discrepancies, unclear areas, or potential inaccuracies in both the summaries and the associated tasks. This thorough validation process, attentive to both content accuracy and the prevention of hallucinations and bias through multiple annotators review, ensures the integrity and quality of our synthetic dataset. Figure 4 in Appendix B presents a screenshot of the annotation platform we used.

Upon further examination of our data, a comparison between machine-generated and human-

Table 1: Comparison of different methodologies for NER. The table showcases various models, their sizes, and the method employed, along with their performance metrics.

Model	Size	Method	F1	Precision	Recall
nlpaueb/legal-bert-small-uncased	35M	Fine-tune	$48.90_{\pm 0.39}$	$41.92_{\pm 0.80}$	$58.69_{\pm 0.52}$
distilbert-base-uncased	66M	Fine-tune	$49.71_{\pm 0.83}$	$42.19_{\pm 0.89}$	$60.50_{\pm 0.77}$
bert-base-cased	108M	Fine-tune	$54.80_{\pm 0.64}$	$47.23_{\pm 1.06}$	$65.28_{\pm 1.01}$
bert-base-uncased	109M	Fine-tune	$53.22_{\pm 1.42}$	$45.86_{\pm 1.68}$	$63.42_{\pm 1.11}$
roberta-base	125M	Fine-tune	**$62.69_{\pm 0.69}$**	$56.58_{\pm 1.12}$	**$70.30_{\pm 0.73}$**
nlpaueb/legal-bert-base-uncased	109M	Fine-tune	$57.50_{\pm 0.94}$	$50.34_{\pm 1.26}$	$67.04_{\pm 0.71}$
lexlms/legal-roberta-base	124M	Fine-tune	$59.73_{\pm 2.03}$	$53.11_{\pm 2.27}$	$68.25_{\pm 1.86}$
joelito-legal-english-roberta-base	124M	Fine-tune	$59.01_{\pm 1.74}$	$52.52_{\pm 2.52}$	$67.40_{\pm 0.85}$
lexlms/legal-longformer-base	148M	Fine-tune	$62.30_{\pm 1.76}$	**$56.78_{\pm 2.14}$**	$69.04_{\pm 1.32}$
lexlms/legal-roberta-large	355M	Fine-tune	$50.23_{\pm 28.1}$	$46.07_{\pm 25.8}$	$55.22_{\pm 30.8}$
lexlms/legal-longformer-large	434M	Fine-tune	$37.63_{\pm 34.4}$	$34.26_{\pm 31.3}$	$41.76_{\pm 38.1}$
joelito-legal-english-roberta-large	355M	Fine-tune	$58.92_{\pm 4.28}$	$52.88_{\pm 4.95}$	$66.59_{\pm 3.22}$
Falcon	7B	QLoRA	$1.00_{\pm 0.50}$	$39.50_{\pm 16.8}$	$0.50_{\pm 0.20}$
Llama-2	7B	QLoRA	$16.3_{\pm 4.10}$	$34.10_{\pm 11.1}$	$11.20_{\pm 2.60}$
OpenAI GPT-3.5	175B	Few-shot	$2.77_{\pm 0.12}$	$1.78_{\pm 0.08}$	$6.23_{\pm 0.29}$
OpenAI GPT-4	-	Few-shot	$13.55_{\pm 0.54}$	$8.29_{\pm 0.37}$	$37.1_{\pm 0.99}$

Table 2: Entity-specific F1 score for the best-performing NER model, 'roberta-base'.

LAW	VIOLATION	VIOLATED BY	VIOLATED ON
$77.57_{\pm 1.35}$	$59.06_{\pm 0.55}$	$76.88_{\pm 2.06}$	$62.83_{\pm 2.57}$

authored content revealed significant similarities. This comparison involved analyzing various linguistic and structural features of the texts. Both displayed identical average sentence lengths. Moreover, there was not significant difference between the character count between the generated content and the human-authored text. Additionally, when comparing the POS tags between the real text and the generated text, by averaging the total counts of each tag occurrences, the average difference was found to be 26% and the median was 16%.

A key part of our validation process was the classification task. In this task, three independent annotators had to distinguish between machine-generated and human-written records, a challenge also noted in recent research (Mitchell et al., 2023; Kirchenbauer et al., 2023). Our annotators' goal was to label each record based on its origin: machine-generated or human-written. The annotators achieved an average F1-score of 44.86%. However, their Cohen's Kappa scores, which were 0.0821, 0.2149, and 0.0988, showed only minor agreement among them. This low level of agreement, as indicated by Cohen's Kappa scores, points out the complexity of the task. It also suggests that our machine-generated content closely resembled human writing, making it difficult even for experts to tell them apart. The use of Cohen's Kappa in our study is supported by its well-known effectiveness in binary classification tasks, especially in data annotation scenarios (Wang et al., 2019).

5 Experiments

In this section, we explore several methods to tackle the challenging and realistic setups that we created. More precisely, we analyzed the performance of language models on these setups by conducting three sets of experiments. (1) We evaluated models that are inspired by the BERT architecture through the process of fine-tuning (Sun et al., 2020). (2) We explored LLMs such as Falcon-7B, Llama-2-7B and Llama-2-13B through the process of parameter efficient fine-tuning (Houlsby et al., 2019; Hu et al., 2021). (3) Thanks to their out-of-the-box generalization capabilities, we assessed OpenAI's GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) models.

5.1 Setup

NER Our dataset is categorized by Cause of Action (CoA). CoA refers to a set of facts or legal reasons that justify the right to sue or seek legal remedy in a court of law. Due to the potential overlap and similarities between different CoAs, there's a risk of data leakage when training models. To mitigate this, we adopted a strategy where CoAs present in the training set were excluded from the test set. This ensures that the model is evaluated on entirely distinct CoAs, preventing any inadvertent training on test data.

NLI Our dataset contains news articles across four legal domains. Given the similarities in the legal merits between these domains, there is a potential risk of data leakage related to the legal attributes of the cases. To address this issue, we employed a leave-one-out approach. In this method, we tested each legal domain separately while training the model on the other domains. This 'leave-one-out' method strengthens the model's ability to generalize by ensuring it is evaluated on entirely unseen data, reducing the risk of overfitting by its small size. By exposing the model to a variety of legal domains during training, but withholding one domain for testing, we mimic real-world scenarios where the model will encounter previously unseen data.

5.2 Model Classes

BERT Models In this setting, we assess the effectiveness of transformer-based language models (Vaswani et al., 2017). We fine-tuned RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019) and BERT (Devlin et al., 2018) models. Additionally, we evaluated their legal counterparts, i.e., Legal-BERT (Chalkidis et al., 2020) and Legal-RoBERTa (Chalkidis* et al., 2023). Furthermore, we evaluated models (Mamakas et al., 2022) based on the Longformer architecture (Beltagy et al., 2020). Following this, we also assessed the Legal-English-RoBERTa models, which are specialized versions tailored for legal English (Niklaus et al., 2023). We utilized the AutoModel family classes from the HuggingFace Transformers library to train the models. Each model was trained for 10 epochs with an initial learning rate of $2e-5$. In addition, we used early-stopping to prevent overfitting.

Open-Source LLMs In this setting, we evaluated Falcon (Almazrouei et al., 2023) and Llama2s (Touvron et al., 2023) performance. More precisely, we considered the 7 billion parametric version of Falcon, and 7 and 13 billion versions of Llama2. Following the success of Parameter Efficient Fine-Tuning methodologies for fine-tuning LLMs, we leveraged QLoRA (Dettmers et al., 2023) due to its superior performance over other methods. Figure 8 shows the prompt that we designed to guide the tuning process.

The prompt has two parts: Input and Output. The Input contains the sentence on which NER and NLI have to be performed. The Output contains the format in which the LLM has to predict the entities contained in the sentence. It is important to note that during inference, we prompt the model to generate the required output by only including the Input section.

We employed HuggingFace's AutoModelForCausalLM class for fine-tuning, available under an Apache-2.0 license[2]. Each model underwent training for 20 epochs with an initial learning rate of 2e-4, a QLoRA rank of 64, and a dropout rate of 0.25. We used this configuration across both NER and NLI tasks.

Closed-Source LLMs We evaluate OpenAI's GPT-4 (OpenAI, 2023) and OpenAI's GPT-3.5 (Brown et al., 2020) models for few-shot NER and NLI without any fine-tuning, using the matching production models of August 2023. We use the Langchain[3] client, available under an Apache-2.0 license, with few-shot prompts, as demonstrated in Figure 9. In all experiments, we set the temperature to 0.7 and used 9 random samples from the training dataset as few-shot examples. We employed the same prompts as those used for open-source models and the same evaluation mechanism. Each API call was repeated five times.

6 Results

6.1 NER

Table 1 presents the performance metrics of various models. Interestingly, BERT-based models with fewer parameters outperform LLMs by a significant margin. This disparity in performance is due to the difference in objective functions that the different model classes use. BERT-based models employ the cross-entropy objective function per token, providing a stronger gradient signal. Furthermore, the label space is well constrained by the number of possible entities in our data set. On the other hand, LLMs have been fine-tuned via causal language modeling, wherein the task is to learn the joint probability distribution of all tokens by maximizing the likelihood of the data. The gradient signal in the case of fine-tuning LLMs is not as fine-grained as cross-entropy. This is because the label space, i.e., the number of possibilities to predict the next token from, far exceeds the number of required entities.

Across BERT-based models, we notice interesting trends. First, *roberta-base* model attains the best performances, achieving an F1 score of 62.69% and Recall of 70.3%. Second, the perfor-

[2] https://github.com/huggingface/transformers
[3] https://github.com/langchain-ai/langchain

Table 3: Macro F1 evaluation of various model architectures for the NLI task across different legal entities.

Model	Consumer Protection	Privacy	TCPA	Wage
nlpaueb-legal-bert-small-uncased	$60.8_{\pm 7.1}$	$49.6_{\pm 14.}$	$47.6_{\pm 11.}$	$56.7_{\pm 6.0}$
distilbert-base-uncased	$79.8_{\pm 2.0}$	$53.9_{\pm 13.}$	$72.1_{\pm 9.3}$	$71.2_{\pm 7.3}$
bert-base-cased	$65.5_{\pm 9.2}$	$39.9_{\pm 18.}$	$58.9_{\pm 16.}$	$65.5_{\pm 13.}$
bert-base-uncased	$69.3_{\pm 7.7}$	$36.3_{\pm 16.}$	$69.5_{\pm 7.2}$	$64.0_{\pm 16.}$
roberta-base	$82.9_{\pm 4.5}$	$62.0_{\pm 5.0}$	$69.5_{\pm 31.}$	$69.7_{\pm 29.}$
lexlms-legal-roberta-base	$45.8_{\pm 5.8}$	$27.3_{\pm 7.9}$	$48.6_{\pm 14.}$	$44.4_{\pm 19.}$
joelito-legal-english-roberta-base	$61.6_{\pm 14.2}$	$33.1_{\pm 12.2}$	$55.8_{\pm 9.95}$	$48.6_{\pm 17.9}$
lexlms-legal-longformer-base	$58.3_{\pm 16.}$	$27.8_{\pm 4.6}$	$54.8_{\pm 11.}$	$54.5_{\pm 11.}$
lexlms-legal-roberta-large	$18.1_{\pm 0.7}$	$20.2_{\pm 8.1}$	$15.3_{\pm 1.8}$	$16.6_{\pm 0.0}$
lexlms-legal-longformer-large	$19.2_{\pm 1.3}$	$17.5_{\pm 0.6}$	$25.5_{\pm 24.}$	$26.3_{\pm 21.}$
joelito-legal-english-roberta-large	$16.4_{\pm 3.3}$	$20.2_{\pm 5.8}$	$47.3_{\pm 30.3}$	$27.3_{\pm 23.9}$
Falcon 7B	$\mathbf{87.2_{\pm 3.1}}$	$\mathbf{84.5_{\pm 8.8}}$	$\mathbf{83.9_{\pm 0.9}}$	$68.5_{\pm 11.}$
Llama-2 7B	$47.2_{\pm 5.9}$	$47.8_{\pm 10.}$	$63.5_{\pm 7.3}$	$63.7_{\pm 14.}$
Llama-2 13B	$63.1_{\pm 8.0}$	$75.2_{\pm 6.5}$	$63.9_{\pm 10.}$	$\mathbf{86.5_{\pm 5.6}}$
OpenAI GPT-3.5	$17.8_{\pm 2.6}$	$18.12_{\pm 3.1}$	$15.09_{\pm 1.9}$	$12.91_{\pm 5.4}$
OpenAI GPT-4	$49.83_{\pm 19.}$	$48.44_{\pm 9.4}$	$37.04_{\pm 7.4}$	$52.48_{\pm 11.6}$

mance across all metrics improved as model complexity grew, except for Longformer-based models and joelito-legal-english-roberta-based models.

Focusing on LLMs, we observed that both open-source and close-source models perform poorly on this task. Closer analysis of predictions indicated incorrect B-token prediction in generated text. These errors were propagated to the next predictions, causing the LLMs to misclassify the tokens and place them into incorrect entities.

6.2 NLI

Table 3 shows domain-specific performances across all model classes. In contrary to trends discovered in the NER experiments, in NLI we noticed that LLMs outperform BERT-based models by a very significant margin. Unlike NER, in NLI, LLMs are fine-tuned to predict only one token, i.e., either of *entailed*, *contradict*, and *neutral*. Additionally, the NLI task had only 312 samples, and LLMs learn relatively better in low data situations and generalize well to out-of-distribution (OOD) test data sets (Brown et al., 2020).

Except for domain *Wage*, *Falcon 7B* achieved the highest performance across domains (*Consumer Protection*, *Privacy*, and *TCPA*). *Falcon 7B* attained the highest Macro F1 metric, demonstrating its OOD capabilities. Among BERT-based models, *roberta-base* once again achieved the best performance, similar to NER tasks.

7 Error Analysis

To improve our models and enrich our understanding, we conducted a thorough error analysis of top-performing models across tasks. This analysis identifies their limitations, providing a clear roadmap for future refinements.

7.1 NER

In evaluating our NER model, the entity type "VIOLATION" exhibited the lowest F1 score. This entity is often lengthy and contextually complex, making it a challenging target for accurate identification. We conducted an error analysis on a subset of hard cases to understand the model's limitations.

The errors fall into three categories: truncation errors, context misunderstanding, and incorrect entity identification. For instance, in the sentence *"I've been getting these [VIOLATION] constant calls on my cell phone from some company that won't quit [VIOLATION]."*, the model predicted *"constant calls on"* instead of the actual entity. This truncation error suggests the model captures only the initial segment but fails to include the entire scope. In another example, *"They've been [VIOLATION] failing to disclose that their educational programs were underperforming [VIOLATION]."*, the model predicted *"disclose"*, indicating a context misunderstanding. Notably, when the model completely misses the target, it often predicts a much shorter entity, suggesting a bias towards shorter answers when uncertain.

The model struggles with the "VIOLATION" entity type, particularly with longer and more complex entities. Fine-tuning the model with a diverse, context-rich training set could improve its performance. Existing literature also suggests that NER models often struggle with complex entities (Dai, 2018), underscoring the need for continued research in this area.

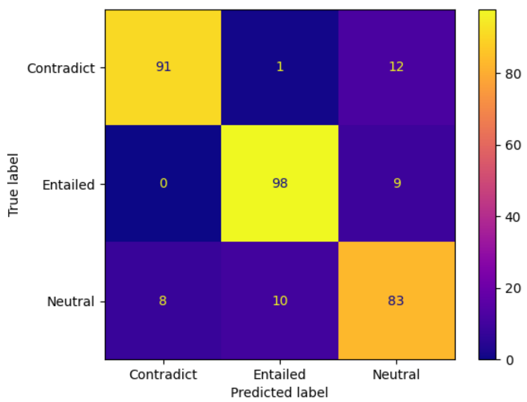

Figure 2: NLI Confusion Matrix derived from the top performer model (Falcon 7B's) predictions.

7.2 NLI

In the error analysis of our best performing NLI model, Falcon 7B, we consolidated the model errors across different legal domains to form a comprehensive view. Our focus was on two types of classification errors: first-class errors, which involve confusions between "Contradict" and "Entailed", and second-class errors, which are misclassifications of "Contradict" or "Entailed" as "Neutral". Figure 2 shows that while Falcon 7B performs well in avoiding first-class errors, it exhibits a substantial number of second-class errors. The high rate of such errors indicates that the model finds it challenging to handle more nuanced cases where it is difficult to discern whether the person was affected by the violation or not.

Although Falcon 7B outperforms other models in this task, it strugglesin accurately classifying statements related to wage areas. This could be attributed to the complexities and ambiguities of the wage norms, which make it challenging to clearly determine whether a wage violation has occurred. Therefore, investigating different token lengths to provide more context or fine-tuning the model to better navigate these intricate wage scenarios could be valuable directions for future work.

8 Conclusions and Future Work

8.1 Answers to the Research Questions

RQ1: *To what extent do our newly introduced datasets enhance the performance of language models in identifying legal violations within unstructured text and associate victims to them?* The study introduced new entities in the datasets. This addition improved the ability of language models to identify legal violations in unstructured text. With these new entities, the roberta-base model achieved an F1-score of 62.69% in identifying violations and 81.02% (Falcon 7B model) in linking them to victims. This demonstrates that our new approach, which focuses on identifying and associating violations to victims, has been successful, yet there remains potential for further refinements and improvements.

RQ2: *How effectively do the language models adapt to new, unseen data for the purpose of identifying legal violations and correlating them with past resolved cases across different legal domains?* Our experiments assessed language models' adaptability to unseen data, especially in the context of identifying legal violations and correlating them with past resolved cases across different legal domains. While BERT-based models demonstrated strong performance in certain tasks, LLMs like Falcon-7B excelled in low-data scenarios, particularly in associating violations with resolved cases. This suggests that these models effectively adapt to new data, especially when the data is limited.

RQ3: *What is the level of difference between machine-generated and human-generated text in the context of legal violation identification?* Our validation process involved a comparison between machine-generated and human-authored content. The findings revealed that the two types of content were strikingly similar in terms of average sentence lengths and character count. When expert annotators were tasked to distinguish between machine-generated and human-written records, they achieved an average F1-score of 44.86%. The low level of agreement among the annotators indicates that our machine-generated content closely resembles human writing, making it challenging even for experts to differentiate between the two.

8.2 Conclusion

In this study, by leveraging LLMs and expert validation, we introduced a dual setup approach to identify legal violations from text. Our approach uses (1) NER to pinpoint violations, resulting in an F1-score of 62. 69% and (2) NLI to associate these violations with resolved cases, resulting in an F1-score of 81.02%. We created two specialized datasets to advance research in this field.

8.3 Future Work

Expanding Legal Areas In future iterations, we aim to expand the dataset to include a broader range of legal areas. By incorporating diverse legal texts, we hope to create a more representative dataset for legal violation identification.

Incorporating Multiple Jurisdictions While our current dataset is heavily focused on common law in US courts, future work will aim to integrate legal texts from various global jurisdictions, including civil law systems. This will not only enhance the datasets diversity but also improve the robustness and applicability of models trained on it.

Fact Matching An avenue for future work is the integration of fact matching. Developing algorithms for cross-referencing facts across sources can enhance the accuracy of legal violation identification, especially when a single source might not provide a complete picture. (Thorne et al., 2018; Jiang et al., 2020)

Limitations

Focus on Common Law in US Courts A primary limitation of our dataset is its focus on US common law. While this deepens understanding of US legal principles and precedents, it may not apply to civil law jurisdictions or non-US legal systems. The nuances, interpretations, and applications of laws can vary significantly across different jurisdictions, and our dataset, being US-centric, might not capture these variations adequately.

Coverage of Areas of Law While our dataset provides a comprehensive overview of legal violations from various text sources, it does have its limitations in terms of the breadth of legal areas covered. The current dataset predominantly focuses on specific areas of law, potentially overlooking nuances and intricacies of other legal domains. For instance, while we have extensively covered areas like consumer protection and privacy, other equally significant areas such as intellectual property, environmental law, or international law might not have been represented with the same depth.

Ethics Statement

The primary objective of this research is to revolutionize the identification and understanding of legal violations within the sprawling landscape of online text. By introducing a novel dataset specifically tailored for Named Entity Recognition (NER) and Natural Language Inference (NLI) tasks in the legal context, we aim to significantly advance the field of Natural Language Processing (NLP) and its applications in law. Our research holds the potential to greatly assist legal professionals in efficiently identifying and addressing legal violations, thereby contributing to a safer and more equitable digital society.

In the pursuit of this objective, we have employed LLMs, specifically GPT-4 (OpenAI, 2023), for data generation, and have subjected the generated data to rigorous validation by expert annotators. This dual-layered approach ensures the quality and reliability of our dataset, while also providing a comprehensive range of examples that can be generalized across various domains.

However, we acknowledge that the deployment of machine learning models in the legal domain is fraught with ethical considerations (Tsarapatsanis and Aletras, 2021). Automating the detection of legal violations could inadvertently lead to false positives or negatives, with serious implications for individual rights and the rule of law. Therefore, we stress that our technology is intended to serve as a supplementary tool for legal professionals, rather than a replacement. It is essential that any application of our dataset and subsequent models be conducted responsibly with a thorough understanding of the limitations and biases that may be inherent in automated systems.

Moreover, we recognize the ethical imperative of data privacy and confidentiality, especially given the sensitive nature of legal texts. All data used in this research have been anonymized and stripped of personally identifiable information to the best of our ability, in compliance with relevant data protection regulations. All the data utilized in this study is sourced from publicly accessible online platforms and does not infringe on any individuals or entities proprietary rights.

Acknowledgements

We express our gratitude to the annotation group at Darrow.ai for their contributions to the dataset.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Orlando Amaral, Muhammad Ilyas Azeem, Sallam Abualhaija, and Lionel C Briand. 2023. Nlp-based automated compliance checking of data processing agreements against gdpr. *IEEE Transactions on Software Engineering*.

Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named entity recognition, linking and generation for greek legislation. In *JURIX*, pages 1–10.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shuyang Cao and Lu Wang. 2022. Time-aware prompting for text generation. *arXiv preprint arXiv:2211.02162*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. *Named entity recognition and resolution in legal text*. Springer.

Xiaodong Gu, Kang Min Yoo, and Sang-Woo Lee. 2021. Response generation with context-aware prompt learning. *arXiv preprint arXiv:2111.02643*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in indian court judgments. *arXiv preprint arXiv:2211.03442*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Daniel Leiker, Sara Finnigan, Ashley Ricker Gyllen, and Mutlu Cukurova. 2023. Prototyping the use of large language models (llms) for adult learning content creation at scale. *arXiv preprint arXiv:2306.01815*.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer.

Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. 2020. A dataset of german legal documents for named entity recognition. *arXiv preprint arXiv:2003.13016*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR 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: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 313–323. Springer.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2023. Anonymity at risk? assessing re-identification capabilities of large language models.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Vasile Păiș, 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.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Amir Saffari, Macro Damonte, and Isabel Groves. 2022a. Clasp: Few-shot cross-lingual data augmentation for semantic parsing. *arXiv preprint arXiv:2210.07074*.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022b. Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. *arXiv preprint arXiv:2209.09900*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023a. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906*.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023b. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

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. *arXiv preprint arXiv:2211.00582*.

Paulo Silva, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, and Marilia Curado. 2020. Using nlp and machine learning to detect data privacy violations. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 972–977.

Stavroula Skylaki, Ali Oskooei, Omar Bari, Nadja Herger, and Zac Kriegman. 2020. Named entity recognition in the legal domain using a pointer generator network. *arXiv preprint arXiv:2012.09936*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash gpt-2 power for event detection. 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 6271–6282.

Juan Wang, Yongyi Yang, and Bin Xia. 2019. A simplified cohen's kappa for use in binary classification data annotation tasks. *IEEE Access*, 7:164386–164397.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yaoquan Yu, Yuefeng Guo, Zhiyuan Zhang, Mengshi Li, Tianyao Ji, Wenhu Tang, and Qinghua Wu. 2020. Intelligent classification and automatic annotation of violations based on neural network language model. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

A Experiments Setting

All experiments were conducted on AWS g5.4xlarge instance, equipped with 1 NVIDIA A10G GPU. The total GPU hours are 85. For each model, the reported metrics are obtained by computing the mean and standard deviation across five runs with randomly initialized weights. All code[4] and datasets (NER[5] and NLI[6]) are available.

A.1 Library Versions

We used the following libraries and associated versions: python 3.8, transformers 4.31.0, seqeval 1.2.2, streamlit 1.25.0, datasets 2.14.2, evaluate 0.4.0, wandb 0.15.7, torch 2.0.1, accelerate 0.21.0, sentencepiece 0.1.99, google cloud aiplatform 1.28.1, openai 0.27.8, langchain 0.0.248, ipython 8.12.2, typer 0.9.0, nltk 3.8, matplotlib 3.7.2.

B Annotation Platform

We ran our annotation platform with the Argilla library [7] available under an Apache-2.0 license.

Figure 4 shows a screenshot of the annotation platform our human experts used.

C Data Distribution

Figure 5 shows the datasets tokens distribution.

Entity	Description	# Labeled Samples
LAW	Specific law or regulation breached.	292
VIOLATION	Content describing the violation.	1326
VIOLATED BY	Entity committing the violation.	292
VIOLATED ON	Victim or affected party.	292

Table 4: Distribution of the NER entities produced by the generation process (2202 in total).

Entity	Description	Labels	# Labeled Samples
Consumer Protection	Deceptive advertising, fraud and unfair business practices.	16/17/29	62
Privacy	Unauthorized collection, use, or disclosure of personal data.	56/54/53	163
TCPA	Unauthorized telemarketing calls, faxes and text messages.	26/27/21	74
Wage	Illegal underpayment and unfair compensation practices by employers.	6/3/4	13

Table 5: Distribution of labeled samples across various legal domains for the NLI task. The number of samples is in the format of Contradiction/Neutral/Entailment.

[4] https://github.com/darrow-labs/LegalLens
[5] https://huggingface.co/datasets/darrow-ai/LegalLensNER
[6] https://huggingface.co/datasets/darrow-ai/LegalLensNLI
[7] https://github.com/argilla-io/argilla

```
You are an human expert who helps generate text based on real-world events.

You should write it in a way human been couldn't detect that it isn't real "platform" text.

Write text which describes how the person was affected and not aware of the lawsuit.

Describe how the person was affected before he even knew about the lawsuit.

The person could be male or female at the age of "age".

Write it "doc type" and "grammar mistakes" .

Don't mention the lawsuit.
Don't mention dates.
Don't mention states.
Don't start with "not allowed words" or any other permutations of those words.
Don't mention money or compensation.

The text should be written as "platform" in "length" "hashtags_emoji".

"agenda"

For example:

Description - Xglasses try-on application used facial recognition to scan the user's face and send it to 3rd parties without the user's consent.

"hypothesis example based on agenda"

event description:
"premise"

The output should be wrap in text tags
<text>
```

Figure 3: Prompt design for generating NLI data set. Prompt contains the task description, specific instructions, and few-shot examples.

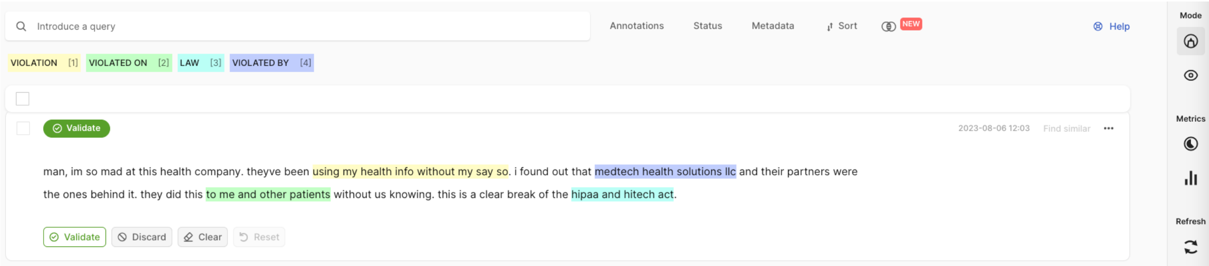

Figure 4: The platform for the human annotations.

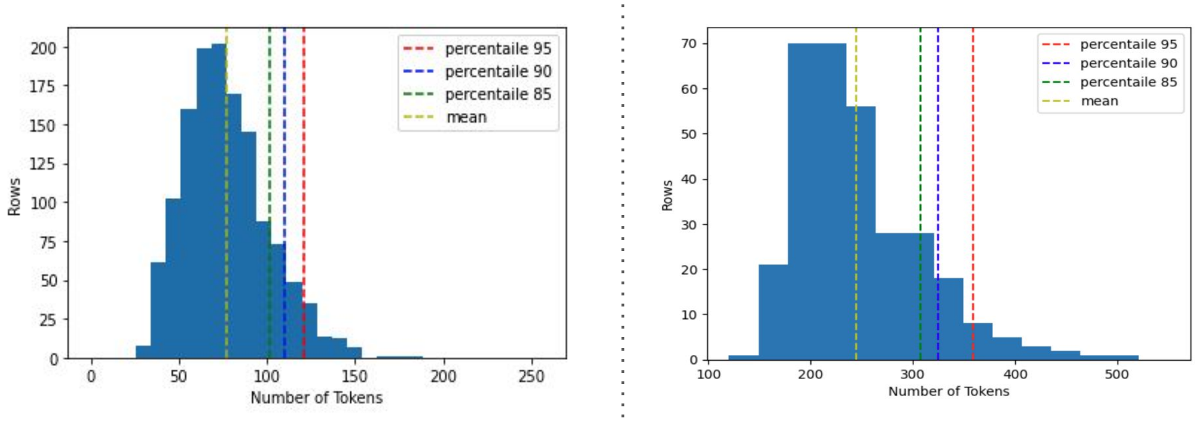

Figure 5: Token Distributions for (left) NER and (right) NLI.

D Prompts

In this appendix, we detail the data generation prompts utilized for the GPT-4 model. The prompts for the datasets creation are illustrated in Figures 6 and 3. Meanwhile, the prompts for fine-tuning can be found in Figure 8. The prompt for the Few-shot approach is depicted in Figure 9

You are an AI assistance that need to write example for training an ml model.
I want you to create for me two examples using this entities:
What is the content that describes the violation. Marks: begin-> [E1], end -> [/E1]

The examples should be paragraphs that contain all of these attributes.

For example:
"I regret to inform potential customers that this banking app has been involved in unsavory practices. They've been caught [E1]quietly charging small unnotified transaction fees and diverting them into undisclosed accounts[/E1]. The inappropriate conduct was led by the company's lead developer and raises serious concerns about the app's credibility."

"The banking app [E1]skimmed undisclosed transaction fees[/E1], led by their lead developer."

The examples should be paragraphs that contain all of these attributes.

Don't stick to the example structure, you can change it as you wish. Use the context below to adjust the story, use augmentation on numbers, dates, names etc to not duplicates examples. You can rephrase the story to other scenarios based on the context. The examples should be from the "coa" cause of action and from the "industry" industry. Don't mention the name of the law in the examples. Write it as a "length" and "text type" text "grammar mistakes" grammar mistakes that has been written as a "doc type". Write each example separately by a newline without numbering prefixes. Don't use any real company/person names. Write it that it will be impossible to know that a model generated this. Context: "context"

(a) Prompt design for Implicit NER data set. Prompt contains the task description, few-shot examples, and specific instructions.

You are an AI assistance that need to write example for training an ml model.
I want you to create for me two examples using these entities:

What is the law that has been broken? Marks: begin-> [E1], end -> [/E1]

What is the content that describes the violation. Marks: begin-> [E2], end -> [/E2]

The violation has been committed by who? This must be explicit and short, don't add non relevant information. Marks: begin-> [E3], end -> [/E3]

The violation has been committed on who (person, group of users etc)? This must be explicit and short, don't add non relevant information. Marks: begin-> [E4], end -> [/E4]

The examples should be paragraphs that contain all of these attributes.

For example:
"The recent case involved a violation of [E1]privacy laws[/E1], where an app was found guilty of [E2]illegally collecting and selling user data[/E2]. It was discovered that [E3]the app developer[/E3] intentionally deceived users by claiming their information would remain secure, but instead, it was being shared with third parties without consent [E4]on unsuspecting users[/E4]."

"In the marketing industry, a prominent advertising agency was found guilty of contravening the [E1]federal trade commission act[/E1] by [E2]misleading consumers with false advertising claims[/E2]. the court determined that [E3]the advertising agency[/E3] had intentionally deceived [E4]the consumers[/E4] by making false claims about the effectiveness of a weight loss product."

"An unsettling incident recently surfaced where an app was indicted for [E2]illegally collecting and selling user data[/E2], constituting a stark violation of [E1]privacy laws[/E1]. Detailed investigations revealed that [E3]the app developer[/E3] had been craftily exploiting [E4]unsuspecting users[/E4], falsely assuring them of data security, whilst secretly passing on their information to third parties."

"Under scrutiny in the realm of marketing was an advertising agency, called to account for [E2]misleading consumers with false advertising claims[/E2]. This breach conspicuously infringed the [E1]federal trade commission act[/E1]. It was adjudicated that [E3]the advertising agency[/E3] had willfully duped [E4]the consumers[/E4] by propagating baseless claims about the efficacy of a weight loss product."

Entities order should be: "entities order". Don't stick to the example structure, you can change it as you wish. Shuffle the appearance of the entities. Use the context below to adjust the story, use augmentation on numbers, dates, names etc to not duplicates examples. You can rephrase the story to other scenarios based on the context. The examples should be from the "coa" cause of action and from the "industry" industry. Write it as a "length" and "text type" text "grammar mistakes" grammar mistakes that has been written as a "doc type". Write each example separately by a newline without numbering prefixes. Don't use any real company/person names. Write it that it will be impossible to know that a model generated this. Context: "context"

(b) Prompt design for Explicit NER data set. Prompt contains the task description, few-shot examples, and specific instructions.

Figure 6: The prompts used for generating the NER data set.

You are tasked with writing examples for training an ML model.

The examples should include entities for the law that has been broken ([E1]), the content that describes the violation ([E2]), the violator ([E3]), and the victim ([E4]).

Each example should be a paragraph that contains all of these attributes.

Use context to adjust the story, use augmentation on numbers, dates, names etc to not duplicate examples.

You are tasked with writing examples for training an ML model.

The examples should include entities for the law that has been broken ([E1]), the content that describes the violation ([E2]), the violator ([E3]), and the victim ([E4]).

Each example should be a paragraph that contains all of these attributes. Use context to adjust the story, use augmentation on numbers, dates, names etc to not duplicate examples. Don't stick to the example structure; you can change it as you wish. Shuffle the appearance of the entities.

You are tasked with writing examples for training an ML model.

The examples should include entities for the law that has been broken ([E1]), the content that describes the violation ([E2]), the violator ([E3]), and the victim ([E4]).

Each example should be a paragraph that contains all of these attributes. Use context to adjust the story, use augmentation on numbers, dates, names etc to not duplicate examples. Don't stick to the example structure; you can change it as you wish. Shuffle the appearance of the entities.

For example: 'The recent case involved a violation of [E1]privacy laws[/E1], where an app was found guilty of [E2]illegally collecting and selling user data[/E2]. It was discovered that [E3]the app developer[/E3] intentionally deceived users by claiming their information would remain secure, but instead, it was being shared with third parties without consent [E4]on unsuspecting users[/E4].

Figure 7: (Top to Bottom) Iterations of prompt design for generating the Explicit NER data set. Prompts contain the task description, and few-shot examples. Figure 6-b contains the final version of prompt used.

INPUT: in the entertainment industry , a significant case has emerged where a company was found guilty of breaking the tcpa (telephone consumer protection act) . the company was found to have repeatedly sent unsolicited promotional emails about concert tickets to consumers , despite their requests to unsubscribe . the court ruled that the company had knowingly violated the law by continuing to send these emails without the express consent of the consumers . despite the consumers numerous attempts to unsubscribe , the company continued its relentless email campaign. \n### OUTPUT: [{O:in the entertainment industry , a significant case has emerged where a company was found guilty of breaking the}, {B-LAW:tcpa}, {I-LAW:(telephone consumer protection act)}, {O:. the company was found to have}, {B-VIOLATION: repeatedly}, {I-VIOLATION:sent unsolicited promotional emails about concert tickets to consumers , despite their request to unsubscribe}, {O:. the court ruled that}, {B-VIOLATED BY:the}, {I-VIOLATED BY:company}, {O:had knowingly violated the law by continuing to send these emails without the express consent of}, {B-VIOLATED ON:the}, {I-VIOLATED ON:consumers}, {O:. despite the consumers numerous attempts to unsubscribe , the company continued its relentless email campaign .}]\n\n

INPUT: in a shocking revelation , it has been discovered that a popular gaming platform has been distributing pirated copies of video games without obtaining the necessary permissions from the original game developers . this act of unauthorized distribution , even in the face of cease and desist letters , has raised serious concerns about the platforms ethical standards . \n### OUTPUT:

(a) Prompt design for NER. (Top) Training prompt, containing the input and output tags, input text, output text and corresponding NER tags. (Bottom) Inference prompt, containing only the input and output tags, input text.

Premise: <Premise text> ### Hypothesis: <Hypothesis text> ### Label: <entailed / contradict / neutral>

Premise: <Premise text> ### Hypothesis: <Hypothesis text> ### Label:

(b) Prompt design for NLI. (Top) Training prompt, containing the input and output tags, premise and hypothesis texts, and corresponding labels. (Bottom) Inference prompt, containing relevant tags, and premise and hypothesis texts.

Figure 8: The prompts used for fine-tuning open-source LLMs across (a) NER and (b) NLI tasks.

You're an AI language model and your task is to perform Named Entity Recognition (NER) on the provided sentence. Label each word in the sentence with the appropriate class based on the context. Use the following classes for labelling:

LAW: This class refers to a law, regulation, act, or any legal entity.
VIOLATION: This class refers to content that indicates a violation of law, a breach of contract, or misconduct.
VIOLATED BY: This class refers to the person, entity or organization that commits the violation.
VIOLATED ON: This class refers to the person, entity or organization that the violation is committed against.

examples:
{examples}

input:
{input}

Given an input consisting of a premise and a hypothesis, determine if the hypothesis supports, contradicts, or is neutral to the premise. The possible labels are: "**Support**", "**Contradict**", and "**Neutral**".

examples:
{examples}

input:
{input}

Figure 9: Few-shot prompt designs for (top) NER and (below) NLI experiments using OpenAI GPT models. Prompts contain input, general task-specific instructions, labels for each task and few-shot examples.