

# LegalLens Shared Task 2024: Legal Violation Identification in Unstructured Text

Ben Hagag\* Liav Harpaz\* Gil Semo\* Dor Bernsohn\* Rohit Saha†  
Pashootan Vaezipoor† Kyryl Truskovskiy‡ Gerasimos Spanakisϕ

\*Darrow AI Ltd., Tel Aviv, Israel {firstname.lastname}@darrow.ai

†Georgian.io, Toronto, Canada {firstname}@georgian.io

‡Scoreinforce, Toronto, Canada {firstname}@Scoreinforce.com

ϕMaastricht University, Maastricht, Netherlands jerry.spanakis@maastrichtuniversity.nl

## Abstract

This paper presents the results of the LegalLens Shared Task, focusing on detecting legal violations within text in the wild across two sub-tasks: LegalLens-NER for identifying legal violation entities and LegalLens-NLI for associating these violations with relevant legal contexts and affected individuals. Using an enhanced LegalLens dataset covering labor, privacy, and consumer protection domains, 38 teams participated in the task. Our analysis reveals that while a mix of approaches was used, the top-performing teams in both tasks consistently relied on fine-tuning pre-trained language models, outperforming legal-specific models and few-shot methods. The top-performing team achieved a 7.11% improvement in NER over the baseline, while NLI saw a more marginal improvement of 5.7%. Despite these gains, the complexity of legal texts leaves room for further advancements.

## 1 Introduction

Legal violations are everywhere, but often go unnoticed. In many areas such as privacy, consumer protection, environmental law, and labor regulations, traces of these violations, indicating wrongdoing, are frequently lost in the vast amounts of digital information. As the world becomes increasingly digital, it is inevitable that traces of these legal violations can be found online. This concept is the foundation of the LegalLens project (Bernsohn et al., 2024). These violations pose significant risks to individuals and institutions, undermining legal and ethical standards in our increasingly digital society. Therefore, developing advanced methods to detect and address these violations is crucial.

Identifying legal violations on the open web presents two primary challenges: first, determining where to search, and second, accurately interpreting whether the information indicates a legal violation. The first challenge involves going

through massive amounts of online content, selecting sources that are likely to yield relevant information while accounting for varying levels of credibility and relevance. The second challenge lies in applying legal knowledge and to determine the legal grounds for these potential violations, and identify victims who may be entitled to compensation.

To advance this field, and to address these challenges, LegalLens tasks were presented in (Bernsohn et al., 2024). The underlying assumption of LegalLens is that Legal violations often leave digital traces, which can be uncovered through careful analysis. LegalLens presented a two-step approach to tackle these challenges: The first is LegalLens-NER (Named Entity Recognition) to extract legal violation entities from online data. The task involves detection and categorization of specific legal violation entities such as laws, violations, violators, and victims within unstructured text. Simple NER methods do not focus on these types of entities and fail to capture the ambiguity of legal language. Figure 3 shows an example of the NER task.

The second step is the LegalLens-NLI (Natural Language Inference) to associate identified violations with relevant legal cases or statutes. More specifically, the task given a premise (allegation summary of a legal case) determine the relationship to a hypothesis (a potential detected violation) and classify their relationship as entailment, contradiction, or neutrality. Figure 2 shows an example of the NLI task.

The datasets for these two sub-tasks were built upon proprietary data by Darrow.ai<sup>1</sup>, designed to be as realistic as possible and to capture the nuances and variability of real-world cases. The data was generated in utilizing GPT-4o (OpenAI, 2023) and domain experts, ensuring both realism and com-

<sup>1</sup><https://www.darrow.ai/>

plexity.

The 1st Shared Task on LegalLens was organized to encourage new research at the intersection of natural language processing and legal studies and to stimulate interest in legal violation detection within the NLP community.

In this paper, we present the results of the Shared Task, offering a detailed description of the evaluation data and the systems developed by participants. We analyze the performance of the participating systems, evaluating their capabilities in processing legal language and identifying legal violations. The top-performing systems for NER showed a substantial improvement over the baseline, with a 7.11% increase in F1 score for the best team. The NLI task saw more marginal progress, with only one team outperforming the baseline by 5.7%. While these improvements highlight progress in legal violation detection, particularly in entity recognition, there remains significant room for further advancements in handling the complexities of natural legal language inference.

As a result, this shared task holds value not just for experts in Machine Learning and NLP, but also for legal professionals, sociologists, and policymakers. This initiative has the potential to foster interdisciplinary collaborations and contribute to advancements in detecting legal violations in the digital era. We are happy to see interdisciplinary teams with participants from CS and NLP alongside legal practitioners, students and social science.

The remainder of this paper is structured as follows: In Section 2, we provide an overview of the LegalLens tasks. Section 3 describes our data collection process, while Section 4 presents the systems and results. Section 5 delves into the details of the three winning teams, and Section 6 offers an overview of the current research landscape.

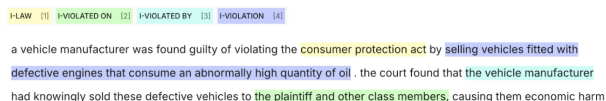


Figure 1 shows a sentence with several entities highlighted in colored boxes: 'LAW' (green), 'VIOLATED ON' (orange), 'VIOLATED BY' (purple), and 'VIOLATION' (blue). The sentence is: "a vehicle manufacturer was found guilty of violating the consumer protection act by selling vehicles fitted with defective engines that consume an abnormally high quantity of oil. the court found that the vehicle manufacturer had knowingly sold these defective vehicles to the plaintiff and other class members, causing them economic harm".

Figure 1: NER sub-task example showing highlighted legal violation entities, including Law, Violation, Violation By, and Violation On.

## 2 What is LegalLens

To efficiently detect legal violations across various domains in the online digital data, a system must be developed that can scan large datasets, isolate

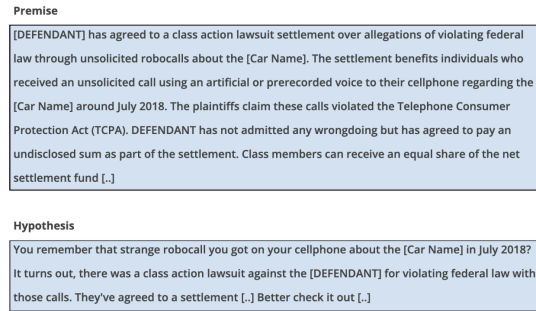


Figure 2 shows an example of the LegalLens NLI task. It consists of two parts: 'Premise' and 'Hypothesis'. The premise is a paragraph about a class action lawsuit settlement over allegations of violating federal law through unsolicited robocalls. The hypothesis is a question: "You remember that strange robocall you got on your cellphone about the [Car Name] in July 2018? It turns out, there was a class action lawsuit against the [DEFENDANT] for violating federal law with those calls. They've agreed to a settlement [...] Better check it out [...]"

Figure 2: An example of the LegalLens NLI task, where the model assesses whether the provided hypothesis (a potential legal violation) is supported, contradicted, or unrelated to the premise (an allegation summary).

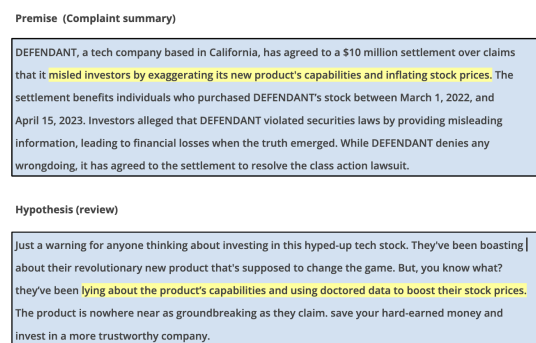


Figure 3 shows an example of the NLI sub-task. It consists of two parts: 'Premise (Complaint summary)' and 'Hypothesis (review)'. The premise is a paragraph about a tech company based in California, DEFENDANT, which has agreed to a \$10 million settlement over claims that it misled investors by exaggerating its new product's capabilities and inflating stock prices. The hypothesis is a review: "Just a warning for anyone thinking about investing in this hyped-up tech stock. They've been boasting about their revolutionary new product that's supposed to change the game. But, you know what? they've been lying about the product's capabilities and using doctored data to boost their stock prices. The product is nowhere near as groundbreaking as they claim. save your hard-earned money and invest in a more trustworthy company."

Figure 3: Example of the NLI sub-task showing how premises like court-filed complaints or articles are used to identify individuals harmed by violations. Both the premise and hypothesis were selected due to matching violation entities identified by the LegalLens-NER model, illustrating the system's ability to link legal grounds to personal experiences and recognize potential victims.

relevant information, and accurately map it to appropriate legal grounds. This involves scanning large amount of online data, contextualizing the findings by linking them to specific legal grounds, and clearly explaining potential violations. Additionally, the system must identify the affected individuals or entities who may be entitled to legal recourse, thereby enabling effective enforcement and remediation.

LegalLens is designed to address these challenges by providing a structured approach to detecting legal violations in digital data. It achieves this by breaking down the task into two key components: LegalLens-NER to identify relevant legal entities and LegalLens-NLI to determine the relationship between data points and legal grounds. In the following section, we will delve deeper into each of these sub-tasks, explaining how they con-

tribute to the overall goal of efficiently detecting and contextualizing legal violations. For the full description of both tasks please refer to the original LegalLens work (Bernsohn et al., 2024).

## 2.1 LegalLens-NER

The LegalLens-NER task in the LegalLens framework aims to identify key legal entities relevant to detecting violations in unstructured text. The LegalLens-NER model classifies tokens into predefined categories: *Law* (the specific law violated), *Violation* (the nature of the infraction), *Violated By* (the responsible entity), and *Violated On* (the affected party).

Given the sheer volume of data and the challenge of identifying relevant information, the LegalLens-NER task acts as an initial filter, extracting critical entities like laws and violations while discarding irrelevant or non-essential information. This process ensures that only the most pertinent data is selected for further analysis and association with legal grounds in subsequent tasks (LegalLens-NLI). The primary goal is to highlight relevant data points for deeper legal examination, making the subsequent steps more efficient and focused.

The dataset for the LegalLens-NER task was curated from class action complaints, with the key violation sections extracted and then summarized and refined using GPT-4. The generated text was formatted as articles, reviews, and social media posts. Human experts validated the realism of the output and annotated the entities. Two prompting strategies were used: explicit, focusing on multiple entities with specific structure, and implicit, centering on a single entity, particularly the violation content. Additional parameters like cause of action and industry were also included to tailor the content to various and real-world scenarios.

## 2.2 LegalLens-NLI

The NLI task in the LegalLens framework is designed to map identified legal violations to the most relevant Cause of Action (CoA) or legal statute. By proceeding the LegalLens-NER model, this model aligns the relevant data with specific legal frameworks, such as laws or precedents, and provides a justification or reasoning for the connection between the identified violations and the applicable legal grounds.

As we worked on this task, we understood that LegalLens-NLI can serve another important pur-

pose: identifying individuals who may have been harmed by the violation. By using descriptions of violations—such as court-filed complaints, articles, or other texts—as premises, the NLI task can analyze relevant online content, like reviews or posts, where people describe their personal experiences. This allows the system to link the identified legal violations to specific individuals who have suffered harm, thus expanding its capability to both identify legal grounds and recognize victims of these violations.

This combined approach strengthens the process of tracing violations back to real-world consequences, making it possible to identify affected individuals with greater precision and relevance.

The dataset is derived from curated legal news articles, with the key violation sections summarized and refined using LLM. The premise for each sample consists of these curated and summarized violation descriptions. The hypotheses were generated using different LLM setups to simulate various scenarios and complaints to reflect real-world situations.

The dataset is labeled with three categories: *Entailment* where the violation is directly supported by the legal grounds; *Contradiction*, where the violation contradicts the legal grounds; and *Neutral*, where the relationship between the violation and the legal grounds is ambiguous or irrelevant.

Human experts validated the correctness and completeness of the premises and hypotheses, and annotated the NLI labels accordingly.

## 3 Dataset Curation for LegalLens Shared Task

Building upon the original LegalLens dataset and addressing its limitations, we have created a more comprehensive and challenging benchmark for LegalLens sub-tasks.

Our goal was to create a dataset that not only mimics real-world scenarios but also presents a challenging benchmark for state-of-the-art NLP models in the legal domain.

The resulting dataset for the shared task maintains the dual-task structure of the original LegalLens, focusing on NER for violation identification and NLI for matching violations with known cases. With improved prompt practices, better annotators guidelines, human expert practices, and following feedback from the original paper (Bernsohn et al., 2024), we improve the gen-

eration process and the resulting annotations, yielding more realistic content, improve data quality and reduce bias.

Our enhancement process consisted of three primary steps: The first was to clean the dataset from duplicated or almost duplicates examples. In some cases we have found that similar patterns appears in the dataset too often that is not making sense. We have tried to detect these patterns and exclude instances that repeat them. Also, to prevent potential biases and ensure broader applicability, we masked all company names within the dataset, including Defendants and Plaintiff names in the NLI dataset. We found that models were prone to overfitting if this masking process was not applied.

Additionally, we have implemented an improved three-step validation where legal experts conducted a multi-stage validation process, including a review for factual accuracy and legal relevance, cross-validation of NER annotations, and examination of premise-hypothesis pairs for logical consistency, completeness and correctness in the NLI task. All annotation conducted via Argilla (Daniel and Francisco, 2023) available under an Apache-2.0 license.

Table 1 shows the datasets tokens distribution. Also, table 2 shows the distribution of labeled samples across various legal domains for the NLI task, formatted as Contradiction/Entailment/Neutral.

Entity	Description	# Labeled Samples
LAW	Specific law or regulation breached.	373
VIOLATION	Content describing the violation.	1665
VIOLATED BY	Entity committing the violation.	373
VIOLATED ON	Victim or affected party.	373

Table 1: Distribution of NER entities generated through the combined dataset from the original paper and the updated process, excluding duplicates.

## 4 System Descriptions and Performance

The competition was hosted on CodaBench<sup>2</sup> (Xu et al., 2022). During the evaluation phase, the leaderboard was hidden, meaning participants did not receive feedback on their submission scores until the phase concluded. Each team was allowed one submission per sub-task.

Both sub-tasks were evaluated as in the original

<sup>2</sup>LegalLens shared task website: <https://www.codabench.org/competitions/3052/>

Entity	Description	Labels	# Labeled Samples
Consumer Protection	Deceptive advertising, fraud and unfair business practices.	28/47/32	107
Privacy	Unauthorized collection, use, or disclosure of personal data.	80/72/82	234
TCPA	Unauthorized telemarketing calls, faxes and text messages.	38/34/39	111
Wage	Illegal underpayment and unfair compensation practices by employers.	9/7/5	21

Table 2: Distribution of labeled samples across various legal domains for the NLI task, formatted as Contradiction/Entailment/Neutral. This dataset combines samples from the original paper and the updated process, excluding duplicates.

paper: the LegalLens-NER sub-task was assessed using the weighted F1 score, to account for class imbalance, with each class’s F1 score weighted by the number of true instances. Evaluation was conducted using the seqeval(Nakayama, 2018) method, which requires exact matches between predicted and true entity spans—both the boundaries and the entity type must match precisely. We followed the IBO format (Inside, Beginning, Outside), where a correct match requires both the boundaries and tags to be accurate. The LegalLens-NLI sub-task used the standard macro F1 score. Participants received the hidden test set only two days before the submission deadline, after submitting the source code of their best architecture. Changes to the model were not permitted after the release of the hidden test set. During the evaluation phase, organizers verified that the predictions could be reproduced using the submitted source code.

### 4.1 Baseline Systems

As a baseline for the each sub-task, we use the best models from the original LegalLens paper (Bernsohn et al., 2024). We trained and evaluated the best models on the new datasets generated for the shared task, as described above. That is to make sure our baseline is up-to-date and performance improvement by participants is by better models, not just by our new dataset. For LegalLens-NER the best model is RoBERTa-base which was fine tuned on the LegalLens-NER dataset. The macro F1-Score for this model is 38.1%. For LegalLens-NLI: the best model is Falcon-7B (Almazrouei et al., 2023) which achieved the highest score of 80.7% macro F1 on average across domains.



## 4.2 Participating teams

A total of 87 individual users grouped in 38 teams participated in the shared task, out of which the highest seven teams elected to write a system description paper. Most of the teams participated in both sub-tasks. Table 3 presents the results for the top six teams in the LegalLens-NER sub-task, Table 4 shows the results for the LegalLens-NLI sub-task, and Table 5 shows an entity level performance for the LegalLens-NER sub-task. Most teams achieved better results than our baseline. Another point worth noting is that success in one sub-task does not necessarily translate to success in the other. Out of the 38 teams, only the NowJ team made it to the top three systems in both tasks. This highlights that the challenges posed by the LegalLens-NER and LegalLens-NLI sub-tasks are distinct, requiring different approaches and strengths.

Lastly, we note that there is a ceiling in terms of performances in the NER task. The top 4 teams achieve score around 70% F1 score, which seems to be the plateau. suggesting that there is room for improvement.

We present the leaderboard for both NER and NLI tasks, showcasing the top six teams and their F1 scores. The next section delves into the leading approaches in each task.

Team Name	Test F1 Score
Nowj	0.416
Flawless Lawgic	0.402
UOttawa	0.402
Baseline	0.381
Masala-chai	0.380
UMLaw&TechLab	0.321
Bonafide	0.305

Table 3: Top six teams for the LegalLens-NER sub-task, with performance measured by weighted F1 scores on a hidden test set.

In the NLI task, the leading team employed a Mixture-of-Experts approach (Jiang et al., 2024), which significantly outperformed the subsequent teams.

All submitted models are available in Darrow.ai’s Hugging Face Space<sup>3</sup>.

<sup>3</sup><https://huggingface.co/darrow-ai>

Team Name	Test F1 Score
1-800-Shared-Tasks	0.853
Baseline	0.807
Semantists	0.785
Nowj	0.746
UOttawa	0.724
bonafide	0.653
masala-chai	0.525

Table 4: Top six teams for the LegalLens-NLI sub-task, with performance measured by Macro F1 scores on a hidden test set.

Team	Law	Violation	V-By	V-On
Nowj	0.7310	0.630	0.041	0.337
Flawless Lawgic	0.711	0.582	0.081	0.310
UOttawa	0.701	0.626	0.045	0.299
Baseline	0.668	0.499	0.087	0.353
Masala-chai	0.636	0.589	0.042	0.308
UMLaw&TechLab	0.596	0.573	0.047	0.104
Bonafide	0.750	0.230	0.152	0.264

Table 5: Entity-specific performance for each team in the LegalLens-NER sub-task, showing F1 scores for the identification of Law, Violation, Violated-By, and Violated-On entities.

## 5 Deeper Analysis

In this section, we describe the key methodologies and innovative techniques employed by the top-performing teams in the LegalLens Shared Task.

### 5.1 LegalLens-NER Methodologies Overview

The NowJ team, which achieved the highest score in the LegalLens-NER sub-task, with 0.416 weighted F1 score, adopted a methodical approach that involved data utilization, preprocessing, and model fine-tuning. They have leveraged both LegalLens-NER datasets, the one from the original paper, and the one introduced for the shared task. The former consisted of 710 training samples and 617 test samples, totaling 1,327 samples. The latter contained 976 samples. To optimize training, the team selected the 976 samples from the LegalLensNER-SharedTask as the training set, with the remaining 351 samples (that are not included in the original dataset) from the LegalLensNER dataset used as the validation set. The model architecture combined a pre-trained language model with a Conditional Random Field (CRF) layer. Pre-trained Language Model - the team used the Legal Longformer (lexlms/legal-longformer-base) (Chalkidis\* et al., 2023), a transformer-based model specialized for legal text. This model produced

contextualized word embeddings, which was used for capturing the semantic nuances of the input text. Conditional Random Field (CRF) Layer modeled dependencies between labels, to ensure valid label sequences by optimizing the Maximum Likelihood Estimate (MLE). The team implemented the forward (Blunsom, 2004) and Viterbi (Forney, 1973) algorithms during training and inference to calculate the probabilities of label sequences and decode the most likely sequence, respectively. Training setting includes: LM: Legal lexlms/legal-longformer-base (Chalkidis\* et al., 2023), Max Sequence Length: 256, Initial Learning Rate: 5e-5, Learning Rate for CRF and Fully Connected Layer: 8e-5, Weight Decay (Fine-Tuning): 1e-5, Weight Decay (CRF and Fully Connected Layer): 5e-6, Batch Size: 16, Total Training Epochs: 30 (Best epoch: 18th), Warmup Proportion: 0.1.

To address the issue of subword tokens in the datasets, where subwords were predicted with the 'X' label, the team implemented a post-processing step. This involved replacing any 'X' label with the label of the preceding token. If the preceding token was a 'B-' (beginning) label, the 'X' label was converted to the corresponding 'I-' (inside) label, ensuring the sequence followed the correct labeling structure.

The uOttawa team, which achieved the third-best score in the LegalLens-NER sub-task, with a 0.402 weighted F1 score, developed their model using the SpaCy library (Honnibal and Montani, 2017). The team implemented preprocessing steps to clean and remove null values and to ensure each token had a corresponding NER tag. The team treated the tokens as features, a transformer model, microsoft/deberta-v3-base for contextual embedding, and a custom NER component via Tok2Vec (Honnibal et al., 2020) layer, to represent tokens in a high-dimensional vector space to capture semantic similarities between words. The model's performance was evaluated after each epoch on a validation set to monitor over-fitting.

## 5.2 LegalLens-NLI Methodologies Overview

The Bonafide team, which achieved the fifth highest score in the LegalLens-NLI subtask, developed a methodology involving data augmentation and model fine-tuning. They used Mixtral 8x7b-instruct-v0.1-hf model (Jiang et al., 2024) to generate paraphrases for both premises and hypotheses across the original 312 rows of data.

The model was prompted to produce realistic rephrasings that retained all the details of the original text, resulting in a final dataset of 936 rows. For model training, the Bonafide team utilized the sileod/deberta-v3-small-tasks-source-nli (Sileo, 2023) encoder, which is based on the DeBERTa-v3-small architecture. This encoder, fine-tuned on tasksource for 250,000 steps and oversampled for long NLI tasks, was further fine-tuned on the augmented dataset. The training dataset was tailored to each legal domain, comprising only synthetic data relevant to that domain, while the test dataset remained unaltered. The hyperparameters used for training included a batch size of 8, a learning rate of 2e-5, and a linear learning rate scheduler. The models were trained for 10 epochs with early stopping to optimize performance. Final predictions on the test dataset were derived by aggregating outcomes from four domain-specific models. The most confident label was selected by calculating the argmax on the confidence levels of all four models.

The 1-800-Shared-Tasks team, which achieved the highest score in the LegalLens-NLI sub-task, with 0.853 macro f1 score, implemented a method involving the use of the FastLanguageModel from the Unsloth library<sup>4</sup>. Their approach focused on fine-tuning the PHI3-Medium-NLI-16bit model, with specific configurations to optimize performance on the NLI task. The model was loaded with a maximum sequence length of 2048 and configured to operate in 4-bit precision to manage computational efficiency. They further enhanced the model using LoRA (Low-Rank Adaptation) adapters (Hu et al., 2021), allowing for the fine-tuning of only 1% to 10% of the model's parameters.

The **NowJ** team, which achieved the third-best score in the LegalLens-NLI sub-task, utilized two datasets provided by the competition organizers on HuggingFace: darrow-ai/LegalLensNLI and darrow-ai/LegalLensNLI-SharedTask. Both datasets contained only a training split with 312 samples. Upon preprocessing, which included converting text to lowercase, removing punctuation, and eliminating extra spaces, they identified approximately 160 differing samples between the two datasets. To maximize data utilization, the participants created a unified dataset comprising the original 312, and the new 160

<sup>4</sup><https://github.com/unslothai/unsloth>

samples. The combined dataset was then split into training and validation sets, with a test size of 0.4, resulting in 283 examples for training (`train_raw`) and 189 examples for validation. Additionally, augmented versions of the examples from the first dataset were appended to create an expanded training set: 665 examples for training set and 189 examples for validation set. The data augmentation implemented using LangChain (Chase, 2022) and the GPT-4o-mini (Achiam et al., 2023) model via API. The goal was to paraphrase both the hypotheses and premises to simulate varying levels of English language proficiency, specifically targeting IELTS<sup>5</sup> levels 6.5 and 8.5<sup>6</sup>. The dataset was expanded with columns to track original and augmented examples, distinguishing versions by IELTS levels. A Pydantic model ensured data consistency, while the GPT-4o-mini model was guided by structured prompts to generate paraphrases. A custom Paraphraser class managed the process, maintaining the integrity of the original meaning. The NowJ team conducted a thorough evaluation of state-of-the-art pre-trained models, including LegalBERT (Chalkidis et al., 2020), T5 (Raffel et al., 2020), and DeBERTa, to identify the optimal architecture for the NLI subtask. DeBERTa (MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli) emerged as the best-performing model due to its stability and high F1-macro scores across multiple training iterations.

Performance across teams varied significantly by legal domain. The 1-800-Shared-Tasks team for instance, performed exceptionally well in structured domains like Privacy (1.0 F1) and TCPA, yet underperformed in complex domains like Wage, likely due to its smaller dataset size and implicit nature of violations. Similarly, NowJ and UOttawa also struggled in domains like Wage, but Semantists fared better due to a more balanced approach across legal domains, highlighting differences in generalization capability across teams. Models fine-tuned on larger datasets showed better overall performance, but those specializing in domain-specific tasks demonstrated marginal improvements, revealing a gap in domain adaptation.

---

<sup>5</sup><https://ielts.org/>

<sup>6</sup><https://ielts.org/take-a-test/preparation-resources/understanding-your-score>

## 6 Related Work

In recent years, there has been increased interest at the intersection of NLP and the Legal domain, with work spanning legal judgment prediction (Chalkidis et al., 2019; Semo et al., 2022; Medvedeva and McBride, 2023) to Information Extraction (Holzenberger and Van Durme, 2023; Bommarito II et al., 2021) to Document analysis (Song et al., 2022; Mamakas et al., 2022) to Text Generation (agarwal-etal-2022-extractive).

In the Information Extraction field, specifically in Named Entity Recognition (NER), (Amaral et al., 2023) have focused on evaluating data agreements for compliance with European privacy laws using NLP techniques. In another study, (Smădu et al., 2022) employed multi-task domain adaptation for NER within the legal domain, showing modest improvements in recall across Romanian and German languages. The work by (Barale et al., 2023) asked language models to detect legal entity types. Additionally, NER has seen increased usage in the legal domain, including efforts to extract entities from court judgment documents in various jurisdictions (Kalamkar et al., 2022). Additionally, (Au et al., 2022) presented E-NER; an Annotated Named Entity Recognition Corpus of Legal Text. However, these entity types, even in legal domain NER tasks, aren't specifically tailored for detecting legal violations and lack the complexity needed for this challenging task. Despite these advancements, existing research typically focuses on a standard set of entity types such as *plaintiff* and *defendant*, with limited exploration of more diverse or nuanced entities relevant to legal violations. Furthermore, these studies are limited in scope, often focusing on specific legal domains or industries.

NLI in the legal domain has gained significant attention in recent years. (Koreeda and Manning, 2021) explored NLI at the document level for contracts, while (Bruno and Roth, 2022) introduced LawngNLI from US legal opinions. (Mathur et al., 2022) presented CaseHoldNLI and a document-level NLI model using optimal evidence selection. (Kwak et al., 2022) introduced a legal NLI dataset for the validity assessment of legal will statements and (Kwak et al., 2023) evaluated the validity of legal will statements across states, using three inputs—statement, condition, and law—to classify the relationship as *support*, *refute*, or *unrelated*. Despite the increased interest, (Bernsohn et al., 2024) is the first to introduce legal violation detection as

a general NLI task across multiple domains.

Prior work has focused on domain-specific use cases, such as privacy protection (Amaral et al., 2023; Silva et al., 2020; Nyffenegger et al., 2023), but these models lack the versatility needed to address the broad spectrum of legal violations across different contexts. LegalLens was the first to establish a cross-domain approach for detecting legal violations.

## 7 Conclusion and Future work

The LegalLens Shared Task demonstrated the potential of leveraging NLP techniques to address the challenge of legal violation detection across diverse domains. Despite the task’s rapid timeline—less than two months from launch to completion—the significant participation of 87 individuals, organized into 38 teams, and the promising results underscore the community’s interest and the relevance of this problem.

We call on the broader research community, particularly those in interdisciplinary fields, to contribute resources, methodologies, and diverse perspectives. Collecting and consolidating these perspectives will deepen our understanding of the complexities within this field. As we refine and build upon the LegalLens framework, we encourage diverse perspectives and innovative approaches that can address the challenges of this important task. Collaboration across disciplines will be crucial in advancing the state of the art in this important area.

The top models achieved a 0.416 F1 score in LegalLens-NER (microsoft/deberta-v3-base) and 0.853 F1 score in LegalLens-NLI (phi3). However, a significant drop was observed in identifying the "Violated By" and "Violated On" entities, indicating room for improvement. This gap suggests the potential for integrating other information extraction techniques, even possibly from outside the legal domain.

Key questions remain unresolved: How will the techniques scale with larger language models and adapt to less-resourced languages? Can we enhance the granularity of legal entity interactions, particularly in more implicit scenarios? Additionally, how will these approaches generalize across broader legal domains and real-world applications?

## Limitations

A challenge of identifying cases of legal violation in the open web is information sparsity. In other

words, these cases do not present themselves in entirety, and in one place. Often times, the salient details of a case are spread across multiple sources on the web, and individually do not offer much insight into the case. It is only when these individual details are stitched together, do they afford themselves to a holistic understanding of the full story, and subsequent evaluation of the case.

## Ethics Statement

We strive to adhere to the [ACL Code of Ethics](#).

Bias and fairness in machine learning have been subjects of long-standing research. As we aim to develop more complex and impactful solutions to address the evolving media and world knowledge, we understand that this goes beyond merely developing or implementing ML algorithms. Inherent biases arise from datasets, task definitions, culture, and even researchers’ beliefs and motivations. Addressing these biases effectively requires collaboration across disciplines. Our technology is designed to supplement, not replace, legal professionals, with responsible application and awareness of potential limitations and biases in automated systems. All data used in this research have been anonymized and stripped of personally identifiable information in compliance with relevant data protection regulations. The data utilized in this study are sourced from publicly available online platforms and do not infringe on any proprietary rights of individuals or entities.

## Acknowledgements

We would like to extend our gratitude to Darrow.ai for providing the dataset, computational resources, and domain expertise that made this research possible. Our thanks also go to the NLLP workshop for facilitating and helping to organize this shared task.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heston, 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*.
- Ting Wai Terence Au, Ingemar J Cox, and Vasileios Lampos. 2022. E-ner—an annotated named entity recognition corpus of legal text. *arXiv preprint arXiv:2212.09306*.
- Claire Barale, Michael Rovatsos, and Nehal Bhuta. 2023. Do language models learn about legal entity types during pretraining? *arXiv preprint arXiv:2310.13092*.
- Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyril Truskovskiy. 2024. Legallens: Leveraging llms for legal violation identification in unstructured text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2129–2145.
- Phil Blunsom. 2004. Hidden markov models. *Lecture notes, August*, 15(18-19):48.
- Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. 2021. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. In *Research handbook on big data law*, pages 216–227. Edward Elgar Publishing.
- 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*.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. [Neural legal judgment prediction in English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.
- 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*, Online. Association for Computational Linguistics.
- Ilias Chalkidis\*, Nicolas Garneau\*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. [LeX-Files 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.
- Harrison Chase. 2022. [LangChain](#).
- Vila-Suero Daniel and Aranda Francisco. 2023. [Argilla - Open-source framework for data-centric NLP](#).
- G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.
- Nils Holzenberger and Benjamin Van Durme. 2023. Connecting symbolic statutory reasoning with legal information extraction. In *Proceedings of the Natural Language Processing Workshop 2023*, pages 113–131. Association for Computational Linguistics.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- 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 Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.
- Alice Kwak, Gaetano Forte, Derek E Bambauer, and Mihai Surdeanu. 2023. Transferring legal natural language inference model from a us state to another: What makes it so hard? In *Proceedings of the Natural Language Processing Workshop*.
- 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.
- Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. [Processing long legal documents with pre-trained transformers: Modding legalbert and longformer](#).

- Puneet Mathur, Gautam Kunapuli, Riyaz Bhat, Manish Shrivastava, Dinesh Manocha, and Maneesh Singh. 2022. Docinfer: Document-level natural language inference using optimal evidence selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 809–824.
- Masha Medvedeva and Pauline McBride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84.
- Hiroki Nakayama. 2018. [seqeval: A python framework for sequence labeling evaluation](https://github.com/chakki-works/seqeval). Software available from <https://github.com/chakki-works/seqeval>.
- 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](#).
- 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.
- 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.
- Damien Sileo. 2023. [tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation](#). *arXiv preprint arXiv:2301.05948*.
- Paulo Silva, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, and Marília 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.
- Răzvan-Alexandru Smădu, Ion-Robert Dinică, Andrei-Marius Avram, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2022. [Legal named entity recognition with multi-task domain adaptation](#). In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 305–321, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Dezhao Song, Andrew Vold, Kanika Madan, and Frank Schilder. 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, 106:101718.
- Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. [Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform](#). *Patterns*, 3(7):100543.