Enhancing Legal Violation Identification with LLMs and Deep Learning Techniques: Achievements in the LegalLens 2024 Competition

Tan-Minh Nguyen[1], **Ngoc-Duy Mai**[1], **Xuan-Bach Le**[1], **Huu-Dung Nguyen**[1]
Cong-Minh Pham[1], **Ha-Thanh Nguyen**[2] and **Thi-Hai-Yen Vuong**[1, *]
[1]VNU University of Engineering and Technology, Hanoi, Vietnam,
[2]National Institute of Informatics, Tokyo, Japan,
[1]{20020081,21020512,22024506,22028076,22028239}@vnu.edu.vn
[2]nguyenhathanh@nii.ac.jp
[*] **Correspondence:** yenvth@vnu.edu.vn

Abstract

LegalLens is a competition organized to encourage advancements in automatically detecting legal violations. This paper presents our solutions for two tasks Legal Named Entity Recognition (L-NER) and Legal Natural Language Inference (L-NLI). Our approach involves fine-tuning BERT-based models, designing methods based on data characteristics, and a novel prompting template for data augmentation using LLMs. As a result, we secured first place in L-NER and third place in L-NLI among thirty-six participants. We also perform error analysis to provide valuable insights and pave the way for future enhancements in legal NLP. Our implementation is available at https://github.com/lxbach10012004/legal-lens/tree/main.

1 Introduction

A violation of law refers to the actions of breaking rules or regulations set by the legal system and authority. These violations harm individuals, organizations, and the principles of fairness and justice, particularly in the digital age. Therefore, developing intelligent systems to detect violations and assist legal experts is essential. Thanks to the exploration of advanced techniques in NLP, prior studies developed specialized models to address the problems of detecting violations automatically (Silva et al., 2020; Yu et al., 2020; Breve et al., 2023). This year, LegalLens (Hagag et al., 2024) is first held with the aim of detecting and monitoring violations in various domains including commercial, privacy, environmental law, and consumer protection. The competition contains two tasks: violation detection via named entity recognition (L-NER) and predicting potential victims of the violation using natural language inference (L-NLI). The L-NER task requires a model to determine four types of entities (law, violation, violated by, violated on) given a passage. The L-NLI task identifies whether the relationship between a complaint (premise) and a review (hypothesis) is entailed, contradicted, or neutral.

The paper reports the work of NOWJ team in both tasks. For the first task, L-NER, independent classification is limited because there are strong dependencies in the output sequence (e.g. B-LAW cannot follow I-LAW, details in Section 3.1). Therefore, we address the problem by following sequence labeling with an architecture of BERT and conditional random field (CRF) to compute output probability jointly. Regarding the second task, one of the main challenges is the lack of a high-quality labeled dataset, whereas general NLI data has been highly developed on large datasets. Thus, we propose a novel prompt for data augmentation using recent LLMs to overcome the shortage of labeled data. State-of-the-art language models are then fine-tuned on augmented training data to develop consistent models for the legal domain.

The following sections of the paper are organized as follows: Section 2 presents prior studies addressing named entity recognition and natural language inference tasks, especially in the legal domain. We describe details of our methodology for two tasks in Sections 3 and 4. Section 5 concludes the paper and points out some future work.

2 Related Work

Legal Named Entity Recognition: NER has been one of the most important tasks in NLP, with various applications in special domains such as biomedicine (Kundeti et al., 2016; Hofer et al., 2018), law (Leitner et al., 2019a; Kalamkar et al., 2022; Au et al., 2022) or cross-domain (Jia et al., 2019). Previously, various classical machine learning methods have been developed to address NER in legal texts such as logistic regression, Support Vector Machines, Naive Bayes, and heuristic-based approaches to extract elements or entities from le-

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 336–345
November 16, 2024 ©2024 Association for Computational Linguistics

gal documents (Chalkidis et al., 2017; Cardellino et al., 2017; Glaser et al., 2018). Another approach addresses NER as a sequence-to-sequence problem and trains a pointer generator network to overcome the absence of noisy training data (Skylaki et al., 2021). Many studies have investigated the performance of transformer-based models, domain-specific embeddings, and neural components (i.e., LSTM, BiLSTM, CNN) combined with CRF (Leitner et al., 2019b; Kalamkar et al., 2022; Keshavarz et al., 2022; Çetindağ et al., 2023), inspired by (Lample et al., 2016). The impacts of CRF, word embeddings, and domain-specific knowledge have proven effective in NER.

Legal Natural Language Inference: NLI, also known as textual entailment recognition has gained interest from researchers in recent years. There are a few law-related resources in NLI, including ContractNLI (Koreeda and Manning, 2021), LawngNLI (Bruno and Roth, 2022), LegalNLI (Yang, 2022), and an annual competition COLIEE (Goebel et al., 2024). However, the cost of constructing high-quality datasets in the legal domain is expensive due to expert-effort requirements in data annotation. Thus, prior studies (Aoki et al., 2022) focused on data augmentation to overcome the limited dataset. Aoki et al. (2022) proposed a data augmentation process based on logical structures of original statutory articles to enrich the training set automatically. Recently, LLMs have shown their state-of-the-art in various NLP tasks, including legal NLP. Nguyen et al. (2024), the winner of the legal statute entailment task in COLIEE 2024, leveraged the powers of LLMs for data augmentation and explore the hidden relations between the premise and hypothesis. Particularly, they summarized the legal article (premise) as complementary information and experimented with various prompting techniques on FlanT5-XXL, an open-source model.

3 Legal Named Entity Recognition

3.1 Problem statement

Given a sequence of tokens $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, the task is to assign a corresponding sequence of labels $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ from a predefined label set $\mathcal{C}$. Our objective is to determine the most likely sequence of labels by maximizing the conditional probability:

$$\hat{y} = \arg\max_{y} P(y \mid x)$$

where $P(y \mid x)$ represents the probability of each label y_i given the token x_i. For the L-NER task, the label set $\mathcal{C}$ utilizes the B-I-O (begin, inside, and outside) tagging scheme and includes 4 entities: *law, violation, violated by,* and *violated on*.

3.2 Data Analysis

We identified two versions of the L-NER datasets. The older version[1] consists of 1327 samples, including a training set with 710 samples and a test set with 617 samples. The newer version[2] contains only a training set with 975 samples. However, we found that the new training set appears in the old data. Therefore, we employ 352 samples of the old data that do not intersect the new data as the validation set. Figure 1 depicts the data sets used for this task.

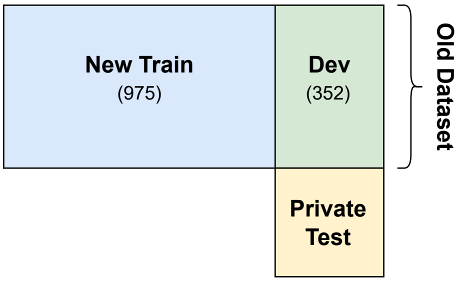


Figure 1: An illustration of our training and validation sets in L-NER.

Table 1 presents the statistics of the training and validation sets in the L-NER task. Further analysis reveals that only 20% samples contain four entities within the sequence. The others include only the *violation* entity type. Additionally, no entity type appears more than once per input sequence. Thus, there is a great imbalance between entities in the dataset. Table 2 shows the statistics of the private test set. The test set contains 380 samples, which is approximately equal to our validation set. Further analysis reveals that the distribution of the private test is quite different from the public data. Firstly, compared to the training and validation sets, the entity distribution is more balanced in the test set. Secondly, the number of *violation* entities in the private test is greater than the number of samples in the test set. While one sequence in the public data only contains as much one time of an entity, a sample in the test set could contain multiple appear-

[1] https://huggingface.co/datasets/darrow-ai/LegalLensNER

[2] https://huggingface.co/datasets/darrow-ai/LegalLensNER-SharedTask

ances of each entity. Finally, the average lengths of entities in the private test set are longer than those in the public set. These differences could pose challenges for models in handling unseen data.

	Training		Validation	
	# Samples	Mean # tokens	# Samples	Mean # tokens
Law	210	3.98	82	4.52
Violation	975	12.30	352	12.57
Violated By	210	2.91	82	3.10
Violated On	210	3.25	82	3.18

Table 1: Statistics of the training and validation sets in L-NER.

	Private Test	
	# Samples	Mean # tokens
Law	246	4.30
Violation	446	16.59
Violated By	399	3.21
Violated On	342	3.66

Table 2: Statistics of the L-NER private test set.

3.3 Methodology

For the L-NER task, we use pre-trained language models combined with a Linear-Chain CRF on top to leverage contextual word embeddings and jointly compute the output probabilities. The architecture is designed to identify and classify named entities within input sequences, as depicted in Figure 2. The vector representation of the input sequence produced by encoders is fed into a linear transformation to map these vectors into a label space. After that, the CRF layer is employed to model the dependencies using these vectors as inputs.

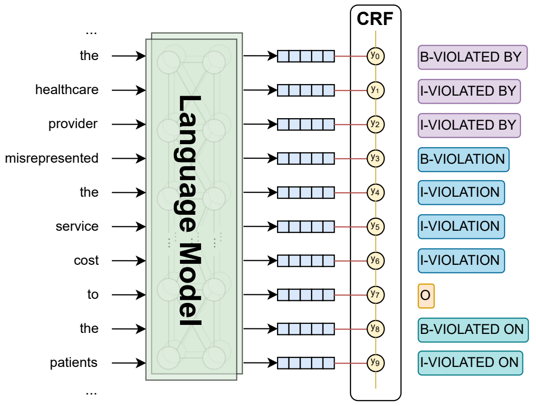


Figure 2: The architecture of BERT-CRF for L-NER task.

3.3.1 Pre-processing

We employ WordPiece (Wu et al., 2016), a subword tokenization technique specifically designed for BERT-based language models. This technique breaks down complex or uncommon words into smaller subword units, enhancing the model's ability to generalize across various word forms. For example, the word "misrepresent" is tokenized into <mis>, <##re>, <##pres>, and <##ent>. After tokenization, the original word labels are realigned with the subword tokens. The first subword retains the original label, while subsequent tokens are assigned a placeholder label (*X*) to ensure label consistency.

3.3.2 Language Model Backbone

Pre-trained language models are utilized to produce contextual embeddings for given input tokens, effectively capturing dependency within the sequence. A linear transformation is then applied to map these embeddings into a label space, with each dimension representing a potential NER tag. This transformation can be represented as follows:

$$\mathbf{H} = Enc(\mathbf{x}) \tag{1}$$

$$\mathbf{P} = \mathbf{H} * \mathbf{W}^{\top} + \mathbf{b} \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^{n \times d}$ is the matrix of hidden states for the token sequence produced by language models, with n being the sequence length and d is the encoder's dimension. $\mathbf{W} \in \mathbb{R}^{k \times d}$ is the weight matrix mapping the hidden dimension d to the number of labels k. $\mathbf{b} \in \mathbb{R}^{k}$ is the bias vector for each label. Finally, $\mathbf{P} \in \mathbb{R}^{n \times k}$ is the emission score matrix, where each row represents a token, and each column represents a label. This sequence of token-level score matrix is then passed to the CRF layer to capture dependencies between labels.

3.3.3 Conditional Random Field

The Linear-Chain CRF is used to model the dependencies between labels in the output sequence. Particularly, CRF assigns a score to each sequence of labels, ensuring that the predicted sequence is globally optimal.

Scoring Algorithm: The score for a sequence of labels $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ given a sequence of input tokens $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is computed as follows:

$$\text{Score}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{n} \left(P_{t,y_t} + A_{y_t,y_{t-1}}\right) \tag{3}$$

where P_{t,y_t} is the emission score for the label y_t at position t, and $A_{y_t,y_{t-1}}$ is the transition score from label y_{t-1} to y_t.

3.3.4 Model Training and Inference

During training, the model parameters are optimized by minimizing the negative log-likelihood loss through backpropagation. Both the LM and CRF layers are trained jointly to maximize the likelihood of the correct label sequences.

Alpha Recursion: The model computes the partition function (normalizing factor) over all possible label sequences. This is expressed as follows:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}'} \exp(\text{Score}(\mathbf{x}, \mathbf{y}')) \tag{4}$$

where the sum is taken over all possible label sequences $\mathbf{y}'$.

Training Objective: The model is trained using the negative log-likelihood (NLL) of the correct label sequence. The NLL loss is given by:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = -\log \frac{\exp(\text{Score}(\mathbf{x}, \mathbf{y}))}{Z(\mathbf{x})} \tag{5}$$

The objective is to minimize this loss, which drives the model to assign higher scores to the correct label sequences.

Viterbi Decoding: During inference, the Viterbi algorithm is applied to decode the most probable sequence of labels for a given input sequence. The decoded labels are then output as the predicted NER tags:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}'} \text{Score}(\mathbf{x}, \mathbf{y}') \tag{6}$$

3.3.5 Post-processing

For pre-existing subwords in the data, which are predicted as *X*, we align them with the label of the preceding token. Only if the preceding token is predicted with a beginning tag (*B-*. . .), the *X* label is converted into an inside tag (*I-*. . .) of the same entity type.

For example, consider the following tokenized sequence and its predicted tags: *[<committed> / O, <against> / O, <mr> / B-VIOLATED ON,* $< . >$ */*$\mathbf{X}$*, <ciesniewski,> / I-VIOLATED ON].* During post-processing, the $\mathbf{X}$ tag for the token $< . >$ is aligned with the preceding *B-VIOLATED ON* tag for $< mr >$ and converted to *I-VIOLATED ON*. This ensures that punctuation or subwords with *X* tags are correctly aligned with the preceding entity labels.

3.4 Experiments and Results

To address the L-NER task, we implement our proposed architecture with different backbone models. This design enables the model to capture both contextual word embeddings from language models and sequential dependencies from the CRF effectively. Following the architecture design, we fine-tune several BERT-based models, including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and Longformer (Beltagy et al., 2020), and compared the performance with their legal-domain counterparts such as LegalBERT (Chalkidis et al., 2020), LegalRoBERTa (Chalkidis* et al., 2023), and LegalLongformer (Chalkidis* et al., 2023). Additionally, we evaluate the BERT-NER[3] model which is a fine-tuned version of BERT for NER tasks. Each model is trained for 30 epochs using the Adam optimizer (Kingma and Ba, 2017), with an initial learning rate of $5e-5$ for the backbone model and $8e-5$ for the CRF layer. All the experiments are carried out on P100 GPU 16GB via the Kaggle platform. We select the best checkpoint on the validation set for each model based on performance metrics. The official evaluation metric for the L-NER task in LegalLens 2024 is the Macro-F1 score, and the results obtained for these models are presented below:

Model	Precision	Recall	F1
BERT-base	0.8675	0.8904	0.8780
Longformer-base	0.8938	0.8861	0.8891
BERT-base-NER	0.8876	0.8925	0.8895
LegalBERT-base	0.8946	0.8907	0.8920
RoBERTa-base	0.8943	0.9002	0.8968
LegalRoBERTa-base	0.9254	0.8939	0.9089
LegalLongformer-base	**0.9264**	**0.9217**	**0.9238**

Table 3: Performances of different backbone models on the validation set.

Table 3 presents the performances of backbone models on the validation set. The model that uses LegalLongformer as the backbone achieves the best scores in all three metrics. There is a slight difference in the performances of the three models BERT, Longformer, RoBERTa. Notably, domain-specific models consistently surpass the general

[3] https://huggingface.co/dslim/bert-base-NER

models by 3.1%pt (i.e. percentage point) to 3.6%pt in precision. This leads to superior performances of three models LegalBERT, LegalRoBERTa, and LegalLongformer in the leaderboard. Our experiments prove the contribution of pre-training language models in specific domains, especially when high-quality data is limited.

According to reports on the validation set, we use the best checkpoint of LegalLongformer-CRF as the submission. The final results of the private test set are presented in Table 4.

Team	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 4: Final leaderboard of L-NER. The top-six teams among thirty-six participants are reported.

Table 4 presents the ranking of the top-six teams on the private test of L-NER. We secured first place in the L-NER task with an F1 score of 0.416, which increases the baseline by 9.1% pt. This result shows the effectiveness of combining CRF and pre-trained language models on the specific-domain NER task. A noteworthy point is the final result is significantly different from the validation result. The baseline method also achieves fourth place in the leaderboard. Indeed, these indicate the challenges of NER in the legal domain. There is room for improving our models' performance and robustness to handle real-world scenarios.

3.5 Error Analysis

Tag	Precision	Recall	F1
B-LAW	0.8870	0.6707	0.7639
I-LAW	0.9299	0.6868	0.7901
B-VIOLATION	0.8138	0.7152	0.7613
I-VIOLATION	0.9021	0.7520	0.8202
B-VIOLATED BY	0.0894	0.0401	0.0553
I-VIOLATED BY	0.1145	0.0572	0.0763
B-VIOLATED ON	0.5106	0.2807	0.3623
I-VIOLATED ON	0.6206	0.2855	0.3911
Macro Average	0.6085	0.4360	0.5027

Table 5: Performance of our model on distinct tags in the L-NER Test Set.

Table 5 shows the performance of our model on different tags in the test set. Overall, our model shows promising results on the *LAW* and *VIOLATION* tags, which capture the violated actions and related law's content. In contrast, identifying two remaining tags is limited, especially with tag *VIOLATED BY*. These two tags capture the entities or organizations in the sequence, one causes the violation, and one is the patient. Further analysis reveals that our model often mistakes the preposition in the tag label. The model also tends to recognize the second occurrence of these entities (person or organizer), while the ground truth labels often pertain to the first occurrence. Furthermore, the length of *VIOLATED ON* tag is relatively short (averaging 3.66 tokens), this pattern negatively impacts the overall performance.

4 Legal Natural Language Inference

4.1 Problem statement

Given an input text pair *(premise, hypothesis)*, the NLI task is to determine the relationship between these texts, whether they are *entailed*, *contradicted*, or *neutral*. This can be framed as a multi-class classification problem, where the goal is to predict the correct category by maximizing the conditional probability of the following:

$$\hat{y} = \arg\max_{y} P(y \mid p, h)$$

Here, p and h denote the premise and hypothesis, respectively. $\hat{y}$ denotes the predicted class, obtained by choosing the class $y \in \{Entailed, Contradict, Neutral\}$ with the highest conditional probability.

4.2 Data Analysis

There are two versions of datasets provided. The older version[4] and newer version[5] both contain 312 samples. After some pre-processing steps, we found that there are 152 samples that both appear in two sets. Therefore, we construct new data consisting of 472 samples, including two public sets, except the intersection part. The train/validation sets are divided with a ratio of $6/4$. The statistics of our dataset for the L-NLI are shown in Table 6. The distribution of labels is uniform, whereas there is no dominant label in the public dataset.

[4] https://huggingface.co/datasets/darrow-ai/LegalLensNLI

[5] https://huggingface.co/datasets/darrow-ai/LegalLensNLI-SharedTask

Label	Samples	Mean # Hypo tokens	Mean # Premise tokens
Contradict	154	71.93	162.14
Entailed	160	75.11	159.87
Neutral	158	62.46	160.99

Table 6: Statistics of the L-NLI public set.

Table 7 presents the statistics of the L-NLI private test set. This test set contains only 84 samples, which is less than five times the size of the public dataset. The private test set exhibits an imbalance, with the *Entailed* label accounting for approximately 50% of the dataset. These differences in the data distribution between the private test and the public sets could negatively impact the model's generalization and consistency.

Label	Samples	Mean # Hypo tokens	Mean # Premise tokens
Contradict	15	43.80	169.46
Entailed	40	59.27	171.05
Neutral	29	40.96	164.62

Table 7: Statistics of L-NLI private test set.

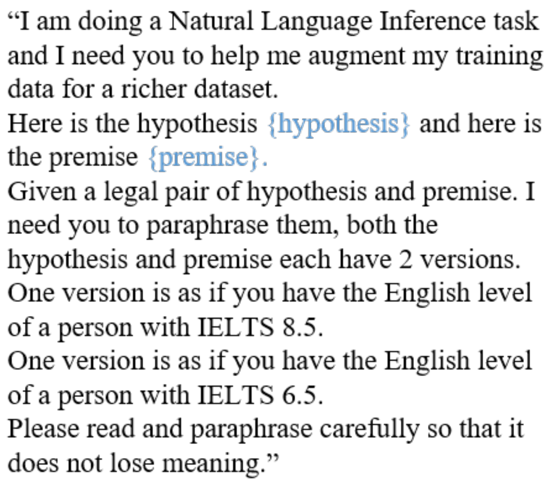
4.3 Methodology

The main difficulty of the L-NLI task is the limited dataset, which consists of 472 samples. Indeed, this would lead to poor generalization and potentially biased outcomes, as the models reflect the narrow perspectives in the datasets. Therefore, we introduce a novel prompt for data augmentation using LLMs. We then fine-tune pre-trained language models on the enriched data to secure stable performances across multiple iterations.

4.3.1 Data Augmentation

To improve the performance and robustness of our models, we employ data augmentation to improve the diversity and variability of the training set. GPT-4o-mini is utilized via the API of OpenAI to generate new data using a novel prompt.

Particularly, we instruct LLMs to paraphrase a hypothesis-premise pair following two styles: one reflecting an IELTS score of 6.5 and the other an 8.5. This approach introduces linguistic diversity in sentence structures, vocabulary, and phrasing while maintaining the core semantic meaning. Figure 3 presents the prompt we used to generate new data. Special symbols {hypothesis} and {premise} are replaced with the content of two paragraphs accordingly. The new training set contains 665 samples including original and augmented data, while the validation set remains the same. New data generated by two levels of IELTS is illustrated in Table 8. Table 9 reports the statistics of the new training set for the L-NLI task.



"I am doing a Natural Language Inference task and I need you to help me augment my training data for a richer dataset.
Here is the hypothesis {hypothesis} and here is the premise {premise}.
Given a legal pair of hypothesis and premise. I need you to paraphrase them, both the hypothesis and premise each have 2 versions.
One version is as if you have the English level of a person with IELTS 8.5.
One version is as if you have the English level of a person with IELTS 6.5.
Please read and paraphrase carefully so that it does not lose meaning."

Figure 3: A novel prompt to generate new data following two styles.

4.3.2 Pre-trained Language Models

We conduct experiments using various language models, including encoder-only and encoder-decoder models.

- **BERT** (Devlin et al., 2018), a widely regarded model, utilizes a multi-layer bidirectional Transformer with an encoder-only architecture. Rather than employing the original base model released by Google, we opted for a domain-specific variant, LEGAL-BERT (Chalkidis et al., 2020), pre-trained on an extensive corpus tailored to the legal domain.
- **T5** (Raffel et al., 2020), on the other hand, employs a full encoder-decoder architecture. This model generates predictions by outputting one of three textual labels: *Entailed*, *Neutral*, or *Contradict*.
- **DeBERTa** (He et al., 2021), while belonging to the BERT family, extends BERT's capabilities by incorporating a novel attention mechanism known as Disentangled Attention, which introduces additional projection matrices to capture relative positional information, integrating them into the attention score computation.

While T5 is fine-tuned following the text generation task, LegalBERT and DeBERTa are fine-tuned

Original data	Been punching in and out with this hand scanner thingy at Superior Health Linens for a while now. You know, the usual 9 to 5 grind. Felt like something out of a sci-fi movie every time I clocked in, but hey, it's the 21st century, right? Technology and all that jazz.
IELTS 6.5	I've been using this hand scanner to clock in and out at Superior Health Linens for a while. It feels like a scene from a sci-fi movie every time I check in, but I guess that's just how things are in the 21st century with all this technology.
IELTS 8.5	I have been clocking in and out with this hand scanner at Superior Health Linens for some time now. Each time I log in, it feels reminiscent of a futuristic film, yet I suppose this is simply the reality of the 21st century, given the prevalence of technology.

Table 8: An example of paraphrasing a hypothesis following different styles.

Label	Samples	Mean # Hypo tokens	Mean # Premise tokens
Contradict	226	66.86	158.08
Entailed	231	72.91	155.81
Neutral	208	58.60	161.53

Table 9: Statistics of the new training set for task L-NLI.

following the sequence classification task. The vector representation of the special token [CLS] is fed into a classification head as follows:

$$H_{cls} = Enc(p, h) \tag{7}$$

$$\mathbf{y} = softmax(H_{cls} * W^\top + b) \tag{8}$$

where $H_{cls} \in \mathbb{R}^d$ is the vector representation of the token [CLS], produced by pre-trained language models, d is the model's hidden size, $W \in \mathbb{R}^{k \times d}$, $b \in \mathbb{R}^k$ are trainable parameters. The output $\mathbf{y} \in \mathbb{R}^k$ represents the predicted probabilities for each class, where $\sum y_i = 1$, k is the number of labels.

4.4 Experiments and Results

Each model is trained for 10 epochs using Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of $5e-6$. The training process is repeated five times to compute the average scores. All the experiments are carried out on P100 GPU 16GB via the Kaggle platform. The official metric for the L-NLI task is the macro F1 score. We experiment with five language models, with and without data augmentation, presented in Table 10 and Table 11. We find that DeBERTa-large outperforms other models in both training cases. Furthermore, DeBERTa models demonstrate stable performance across multiple iterations. The number of parameters also contributes to the results, whereas large models consistently surpass base models. Notably, training models on augmented data could improve the results in all metrics. Particularly, the F1 score saw a rise of 2.9%pt to 5.5%pt on the validation set. Indeed, these results highlight the contribution of data augmentation in handling legal downstream tasks. We select the DeBERTa-large checkpoint with the highest performance on the validation set as the final submission.

Model	Precision	Recall	F1
LegalBERT-base	0.8378	0.8401	0.8342
T5-base	0.8421	0.8676	0.8502
T5-large	0.8685	0.8645	0.8717
DeBERTa-base	**0.8943**	0.8788	0.8831
DeBERTa-large	0.8895	**0.8917**	**0.8848**

Table 10: Average performances of models on the validation set. Before data augmentation.

Model	Precision	Recall	F1
LegalBERT-base	0.8801	0.8813	0.8722
T5-base	0.8882	0.9016	0.8977
T5-large	0.9058	0.9063	0.9043
DeBERTa-base	0.9126	0.9052	0.9089
DeBERTa-large	**0.9210**	**0.9220**	**0.9204**

Table 11: Average performances of models on the validation set. After data augmentation.

Table 12 presents the results of the top six teams in the competition. Our model achieves the F1-macro score of 0.746 on the private test set, placing third place among thirty-six participants. Even though our model could achieve impressive performance on the validation set, it is limited on the private test set. In addition, only the winning team could surpass the baseline in the L-NLI task. Indeed, these results emphasize the challenge of legal downstream tasks.

Team	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 12: The final leaderboard of the L-NLI task.

4.5 Error Analysis

Table 13 presents the error analysis of our model on the L-NLI test set. The proposed method achieves a promising performance on the neutral label, where the precision and recall scores are balanced. In contrast, there is a trade of pattern between results of labels contradict and entailed. While the recall score of the contradict label is 1.0, the model gets a 0.9615 precision score on the entailed label. This result suggests that our model is heavily biased toward the contradict label if there is a relationship between two texts. Another noteworthy point is that approximately 50% of the wrong predictions belong to the Biometric Information Privacy Act (BIPA) domain as shown in Figure 4. This could be attributed to the lack of BIPA area in the training set. Future work could focus on exploiting logical knowledge to reinforce the model's reasoning and inference abilities, which would help to better distinguish the contradict and entailed relations.

	Precision	Recall	F1-score
Contradict	0.5556	1.0000	0.7143
Entailed	0.9615	0.6250	0.7576
Neutral	0.7419	0.7931	0.7667
Macro Average	0.7530	0.8060	0.7462

Table 13: Error analysis of our model on L-NLI test set.

5 Conclusion

This paper presents our work in the LegalLens competition. For the L-NER task, we leverage the contextual embeddings of BERT-based models and compute sequence dependency using a Linear-Chain CRF layer. For the L-NLI tasks, we propose a novel prompt to generate synthesis data using LLMs. The experiments highlight the effectiveness of data augmentation in improving language models' performance. Consequently, we secured first place in L-NER and third place in L-NLI. We also perform error analysis to offer valuable insights and groundwork for future advancements in legal NLP. Future work would focus on improving the robustness and performance of models by exploiting the integration of logical knowledge and LLMs.

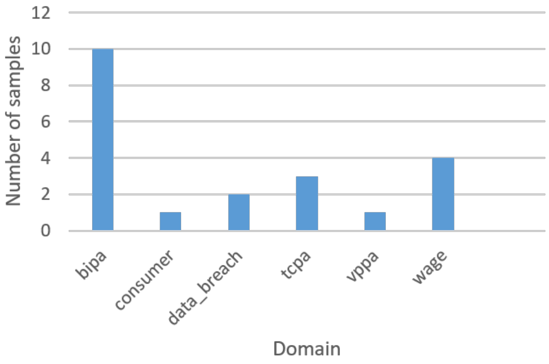


Figure 4: Statistics of wrong prediction's domain.

Limitations

We outline the following limitations in this work: (1) one of the main challenges is the shortage of datasets. Even though we employed data augmentation with LLMs in task 2, the dataset remained limited, affecting the diversity and generalization of our model. Therefore, there is a decline in the performance of our method on the private test set compared to the validation set. Furthermore, the data augmentation using LLMs should be further discussed and studied, to ensure the quality of enriched data. (2) Although domain-specific models are utilized in this work to address legal downstream tasks, the legal logic reasoning is not yet considered explicitly. Indeed, this approach should be studied throughout to enhance the reliability and accuracy of deep learning models in the legal domain. (3) The use of closed-source models like GPT4 is limited by many constraints, which may pose difficulty in reproducing our experiments. Well acknowledging the problem, we would make our code and implementation publicly accessible in the future. Nonetheless, the discussions and insights in this work demonstrate the promising benefits of leveraging LLMs and deep learning techniques for legal violation identification.

Acknowledgments

This work was supported by AIP challenge funding in relation with JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4, by JSPS KAKENHI

Grant Numbers, JP22H00543, JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4 and the MEXT "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project.

References

Yasuhiro Aoki, Masaharu Yoshioka, and Youta Suzuki. 2022. Data-augmentation method for bert-based legal textual entailment systems in coliee statute law task. *The Review of Socionetwork Strategies*, 16(1):175–196.

Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — an annotated named entity recognition corpus of legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Bernardo Breve, Gaetano Cimino, and Vincenzo Deufemia. 2023. Identifying security and privacy violation rules in trigger-action iot platforms with nlp models. *IEEE Internet of Things Journal*, 10(6):5607–5622.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5019–5043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 9–18, New York, NY, USA. Association for Computing Machinery.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 19–28, New York, NY, USA. Association for Computing Machinery.

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

Ilias Chalkidis*, 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.

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.

Ingo Glaser, Bernhard Waltl, and Florian Matthes. 2018. Named entity recognition, extraction, and linking in german legal contracts. In *IRIS: Internationales Rechtsinformatik Symposium*, pages 325–334.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In *New Frontiers in Artificial Intelligence*, pages 109–124, Singapore. Springer Nature Singapore.

Ben Hagag, Liav Harpaz, Gil Semo, Dor Bernsohn, Rohit Saha, Pashootan Vaezipoor, Kyryl Truskovskyi, and Gerasimos Spanakis. 2024. Legallens shared task 2024: Legal violation identification in unstructured text. *Preprint*, arXiv:2410.12064.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. 2018. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

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

Hossein Keshavarz, Zografoula Vagena, Pigi Kouki, Ilias Fountalis, Mehdi Mabrouki, Aziz Belaweid, and Nikolaos Vasiloglou. 2022. Named entity recognition in long documents: An end-to-end case study in the legal domain. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2024–2033.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. 2016. Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1937–1945.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019a. Fine-grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany. Springer. 10/11 September 2019.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019b. Fine-grained named entity recognition in legal documents. In *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham. Springer International Publishing.

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.

Phuong Nguyen, Cong Nguyen, Hiep Nguyen, Minh Nguyen, An Trieu, Dat Nguyen, and Le-Minh Nguyen. 2024. Captain at coliee 2024: Large language model for legal text retrieval and entailment. In *New Frontiers in Artificial Intelligence*, pages 125–139, Singapore. Springer Nature Singapore.

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.

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. 2021. Legal entity extraction using a pointer generator network. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 653–658.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhanye Yang. 2022. Legalnli: natural language inference for legal compliance inspection. In *International Conference on Advanced Algorithms and Neural Networks (AANN 2022)*, volume 12285, pages 144–150. SPIE.

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.

Can Çetindağ, Berkay Yazıcıoğlu, and Aykut Koç. 2023. Named-entity recognition in turkish legal texts. *Natural Language Engineering*, 29(3):615–642.