# Bonafide at LegalLens 2024 Shared Task: Using Lightweight DeBERTa Based Encoder For Legal Violation Detection and Resolution

**Shikha Bordia**
bordiashikha06@gmail.com

## Abstract

In this work, we present two systems—Named Entity Resolution (NER) and Natural Language Inference (NLI)—for detecting legal violations within unstructured textual data and for associating these violations with potentially affected individuals, respectively. Both these systems are lightweight DeBERTa based encoders that outperform the LLM baselines. The proposed NER system achieved an F1 score of 60.01% on Subtask A of the LegalLens challenge, which focuses on identifying violations. The proposed NLI system achieved an F1 score of 84.73% on Subtask B of the LegalLens challenge, which focuses on resolving these violations by matching them with pre-existing legal complaints of class action cases. Our NER system ranked sixth and NLI system ranked fifth on the LegalLens leaderboard. We release the trained models and inference scripts[1].

## 1 Introduction

Social networks and other online platforms are increasingly becoming effective tools to address consumer complaints; however, the vast amount of unstructured textual data makes it challenging to identify valid complaints and if they are associated with any legal violations. There is a pressing need to develop sophisticated methods to identify these hidden breaches, as they have significant implications for individual rights and legal obligations, if any.

In this regard, Bernsohn et al. (2024) propose two subtasks —Subtask A, Legallens NER (Named Entity Recognition), to detect legal violations mentioned in the text and Subtask B, Legallens NLI (Natural Language Inference), to match the detected violations with resolved class action cases. To address these subtasks in this paper, we propose NER and NLI models based on training DeBERTaV3 encoders. We finetune task-specific encoders on our synthetically augmented dataset. In summary, we list our findings here:

1. Continuing to pretrain an already powerful general domain task-specific model on our subtask can boost the performance of our system.

2. While synthetic data can significantly boost the capabilities of models, it's crucial to recognize that surpassing specific thresholds of training data volumes may not necessarily result in proportional enhancements in performance.

3. Scaling laws suggest that Large Language Models (LLMs) show predictable performance improvements. However, smaller models can either match or perform better using appropriate training objectives and data, specifically for classification tasks.

In Section 2, we examine the related works on NER and NLI tasks. Section 3 provides an overview of the methodologies employed for the tasks. In Section 4, we describe the experimental setup. Section 5 discusses the results and findings; Section 6 discusses the conclusions.

## 2 Related Works

**NER** Research in NER has evolved from statistical models such as Maximum Entropy (Borthwick et al., 1998), Hidden Markov Models (Bikel et al., 1999), and Conditional Random Fields (CRF)(McCallum and Li, 2003), using bidirectional RNNs, often combined with CRF layers (Huang et al., 2015; Ma and Hovy, 2016; Lample et al., 2016) to using transformer-based models (Vaswani et al., 2017). This transition has enabled

[1] https://github.com/BordiaS/LegalLens_inference



*Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 259–266

Table 1: The distribution of number of words by entity type in the LegalLens NER training dataset

| LAW | VIOLATED BY | VIOLATED ON | VIOLATION |
|---|---|---|---|
| 4.14 | 2.19 | 3.24 | 12.39 |

accurate and robust entity recognition across various domains and languages. In legal domain, variations of BERT-based transformers (Devlin, 2018) like RoBERTa (Liu, 2019), DeBERTaV3 (He et al., 2021a), LegalBERT (Chalkidis et al., 2020), with BiLSTM and CRF layers on the top (Huo et al., 2023; Ningthoujam et al., 2023) have given state-of-the-art performance on legal NER tasks (Kalamkar et al., 2022; Modi et al., 2023). Legallens NER task has four sets of entity types that have not been previously explored in legal NER research. In this work, we use the recently proposed DeBERTaV3 based GLiNER (Zaratiana et al., 2023) architecture that outperforms both ChatGPT (Brown et al., 2020) and fine-tuned LLMs in zero-shot evaluations on various NER benchmarks.

**NLI** NLI classifies the logical relationship between a premise (a given statement) and a hypothesis (a proposed conclusion) as entailed, contradictory, or neutral. Early work on NLI focused on rule-based systems and logical inference (Giampiccolo et al., 2007). The advent of large-scale datasets, such as the SNLI (Bowman et al., 2015), MultiNLI corpus (Williams et al., 2017), XNLI (Conneau et al., 2018) enabled the development of sophisticated models. Transformer-based models such as RoBERTa (Liu, 2019) and XLNet(Yang, 2019) have pushed the limits of NLI performance by giving human-like scores.

NLI is a critical task in NLP that serves as a benchmark for natural language understanding. Although significant progress has been made, challenges remain in developing systems that can perform robust and generalizable inference across diverse domains and languages. In this work, we use Tasksource's NLI model and finetune it on the LegalLens NLI dataset. Tasksource is a framework that harmonizes data sets for multitask learning and evaluation in NLP by providing a collection of pre-processing methods (Sileo, 2024).

**Data Augmentation** The advent of LLMs has introduced a novel approach to data augmentation in machine learning tasks (He et al., 2021b; Gan and Ng, 2019; Hosseini et al., 2024). Leveraging the capabilities of these models, we employ two distinct strategies to enhance our datasets. For the NER task, we utilize few-shot learning techniques to expand the existing dataset. This method allows us to generate additional, contextually relevant examples based on a small number of initial samples. Concurrently, for the NLI dataset, we implement a paraphrasing approach. This technique involves reformulating the sentences—premise and hypothesis—while preserving their semantic content, thereby increasing the diversity and robustness of our training data. This approach also preserves the original label distribution. We use Mixtral 8x7B model (Jiang et al., 2024), a state-of-the-art LLM, to augment both the datasets. The specific prompts used for these augmentation tasks are detailed in the Appendix A for both the subtasks, ensuring transparency and reproducibility of our methods.

## 3 Methodology

In this section, we introduce our approach for each of the subtasks.

### 3.1 Subtask A: LegalLens NER

**Problem Statement** The NER task aims to detect legal violations in social media posts and online reviews. The training and development datasets consist of 710 and 617 data points. We specifically identify the following entities: `LAW` (law or regulation breached), `VIOLATION` (content describing the violation), `VIOLATED BY`(entity committing the violation) and `VIOLATED ON` (victim or affected party).The average number of words range between 2.19 and 4.14 for `LAW`, `VIOLATED BY` and `VIOLATED ON` while the average number of words for `VIOLATION` is 12.39 as shown in Table 1.

**Contribution** Our main contributions are as follows:

- We finetune a lightweight bidirectional transformer encoder GLiNER proposed by Zaratiana et al. (2023), that uses DeBERTaV3 (He et al., 2021a) as backbone. It is trained on Pile-NER dataset (Zhou et al., 2023).

- We experiment with the architectures—single, bi-encoder and polyencoder—proposed by Zaratiana et al. (2023)

All the pre-trained checkpoints of these models are taken from the Hugging Face hub repository.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| gliner_small-v2.1 | 70.26 | 45.83 | 55.47 |
| gliner_base | 71.30 | 47.02 | 56.67 |
| gliner_small | 72.32 | 45.71 | 56.02 |
| gliner-bi-base-v1.0 | **83.30** | 46.31 | **59.53** |
| gliner-bi-small-v1.0 | 74.00 | 48.39 | 58.52 |
| gliner-poly-small-v1.0 | 71.04 | **49.64** | 58.44 |

Table 2: Comparison of different GLiNER architectures on LegalLens NER development dataset. The table showcases the models and their respective performance

### 3.2 Subtask B: LegalLens NLI

**Problem Statement** The NLI task aims to link resolved class action cases with violations detected by the NER model. The premise comprises summaries of legal news articles, while the hypothesis consists of synthetically generated social media posts that mimic potential legal violations. The dataset includes 312 data points across four legal domains: Consumer Protection, Privacy, TCPA, and Wage.

**Contribution**

- We finetune a multitask DeBERTaV3 based encoder, Tasksource (Sileo, 2024), that casts all the classification tasks as natural language inference and trains the model on 600+ English tasks simultaneously to achieve state-of-the-art performance at its size.
- We propose synthetic data generation to enhance the performance of the model. We employ Mixtral 8x7B by Jiang et al. (2024) to generate paraphrases for each premise-hypothesis pair. The class labels (Entailed, Contradict, and Neutral) remain unchanged. This approach doubles the size of the training data while preserving the original class distribution.
- Augmenting the NLI dataset boosted the final F1 score metric by a significant margin of 7.65%.

## 4 Experimental Settings

**NER** We finetune the GLiNER models on the LegalLens NER dataset using a dropout rate of 0.5 and a batch size of 8. We employ AdamW optimizer with a base learning rate of 1e-5 for pre-trained layers (the transformer backbone, DeBERTaV3) and 5e-5 for non-pre-trained layers (FFN layers, span representation). The model is trained

| Entity Type | Precision | Recall | F1 |
|---|---|---|---|
| LAW | 73.40 | 92.00 | 81.66 |
| VIOLATED BY | 88.16 | 89.33 | 88.74 |
| VIOLATED ON | 71.43 | 73.33 | 72.37 |
| VIOLATION | 68.17 | 39.29 | 49.85 |
| micro avg | 71.93 | 51.49 | **60.01** |

Table 3: Entity level metrics of the best performing model **gliner-bi-base-v1.0** integrated with predefined rules

to a maximum of 10 epochs, starting with a 10% warm-up phase, followed by a decay phase using a linear scheduler. We save the best checkpoint and, subsequently, reduce the learning rate to 5e-6, and train this checkpoint until convergence. To address class imbalance, we use focal loss, instead of cross-entropy loss, with alpha 0.75 and gamma 2.

We experiment with three different architectures proposed by Zaratiana et al. (2023) and Knowledgator Engineering[2]—original GLiNER, the bi-encoder and the poly-encoder as shown in Table 2. During inference, we utilize a model threshold of 0.8 to compute performance metrics. Additionally, we implement a rule to eliminate false positive entities. In the event that multiple entities of the same type are extracted, we discard the entity with the lowest confidence score and retain the one with the highest score. This approach resulted in an improvement in the F1 score by 0.5%, reaching 60.01%.

**NLI** We train four models and test them on each legal domain. Each of these four models is trained on three domains at once and tested on the fourth to prevent data leakage as described by Bernsohn et al. (2024). For each domain, we finetune Tasksource's NLI model using a learning rate of 2e-5, a sequence length of 256, and a batch size of 8 for a maximum of 7 epochs using a cosine scheduler. We

[2]Knowledgator Blog link

| Model | Consumer Protection | Privacy | TCPA | Wage | Macro F1 |
|---|---|---|---|---|---|
| `tasksource` (orignal) | 85.48 | 76.07 | 62.16 | 81.56 | 76.31 |
| `tasksource` (augmented) | **88.71** | **85.88** | **79.72** | **84.61** | **84.73** |

Table 4: Comparison of Tasksource's model performance on LegalLens NLI's dev dataset. The second row shows the improved performance using the augmented dataset over the original dataset as the training data by 7.65%

save the best checkpoint and reduce the learning rate to 2e-6, and further train it until convergence. As shown in Table 4, the synthetically augmented dataset boosted the performance of the models on the development dataset by 7.65%.

## 5 Results and Discussions

**NER** The original GLiNER architecture employs bi-directional encoder. The entity labels, separated by [SEP] token, and the input sequence are concatenated and then passed through the encoder model. The bi-encoder architecture decouples the entity labels and input sequence. The poly-encoder uses fuses the entity label and input sequence encoder representations together to capture the interactions between them. The bi-encoder model, `gliner-bi-base-v1.0`, has best performance with an F1 score of 59.53% and the highest precision of 83.30%. The polyencoder model, `gliner-poly-small-v1.0`, gave the highest recall of 49.64% as shown in Table 2.

Our experiments reveal that shuffling entity order and randomly dropping entities did not affect the metrics. After identifying the best model, we trained it on a synthetic dataset generated using few-shot learning. However, this approach did not yield any improvement in results. We then applied rule-based entity filtering, which improved the development dataset results by 0.5%, increasing the final F1 score from 59.53% to 60.01%. The system ranked sixth on the leaderboad with an F1 score of 33.00% on the test dataset (Hagag et al., 2024).

Table 1 illustrates the distribution of word count by entity type. The `VIOLATION` entity type averages 12.39 words, compared to a maximum of 4.14 for the other three types, increasing the complexity of the task. The model performs better on shorter entities, as shown in Table 3. Previous research has shown that NER models struggle with complex entities and tagging long sequences (Dai, 2018).

Although our model results did not surpass the baselines (Bernsohn et al., 2024), further exploration of medium and large variants of GLiNER could be beneficial. Due to limited computational resources, we were unable to include them in our experiments.

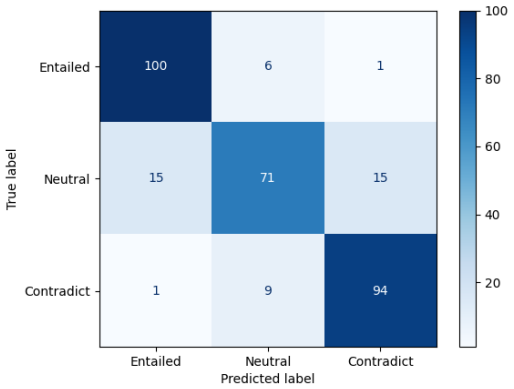


Figure 1: Final Confusion Matrix on the LegalLens NLI the dev dataset

**NLI** For each legal type category, we employ four distinct models. During the evaluation process on the unlabeled test set, we consider the collective assessment of all four models. The final label for a premise-hypothesis pair is determined by the model exhibiting the highest confidence score among the four. Our findings indicate that data augmentation proved beneficial, albeit to a certain extent. When we expanded the dataset to triple its original size by incorporating an additional set of paraphrases, we observed that the corresponding increase in F1 scores was not proportional to the increase in data volume. This suggests that there may be diminishing returns in terms of performance improvement beyond a certain threshold of data augmentation.

We compare our results with the baseline proposed by Bernsohn et al. (2024). They finetune Falcon 7B (Almazrouei et al., 2023) and report an F1 score of 81.02% compared to 84.73% for our model. The system ranked fifth on the leaderboad with an F1 score of 65.30% on the test dataset (Hagag et al., 2024).

In the error analysis of the final model, we see that both the models are capable of handling first class errors—confusions between `Contradict` and `Entailed`. However, our model does better with

handling second-class errors—misclassification of `Contradict` or `Entailed` as `Neutral`; and Falcon 7B model does better with handling another class of errors—misclassification of `Neutral` as `Contradict` or `Entailed`. The confusion matrix for our model is shown in Figure 1.

It is interesting to note that a multitask DeBERTa based encoder surpassed the performance of a 7B parameter by 3.17%. Our model is capable of resolving the ambiguities and complexities related to wage norms. Finally, it can be stated that paraphrasing can serve as a data augmentation technique to enhance the natural language understanding capabilities of smaller models.

## 6 Conclusion

In conclusion, we present two systems developed for the LegalLens 2024 shared task, comprising a zero-shot bidirectional DeBERTa encoder with domain-adaptive pretraining for the NER subtask and a multitask DeBERTa encoder enhanced by data augmentation techniques for the NLI subtask. The experiments demonstrate that synthetic data generation can enrich datasets and improve the performance of encoder-based models. However, it is evident that more data does not necessarily translate to better performance. By optimizing on smaller but richer datasets and employing suitable training objectives, smaller models can outperform larger language models.

Further exploration of different augmentation strategies, with a particular focus on generating more contextually diverse synthetic data, employing adversarial data, or leveraging domain-specific paraphrasing techniques, may yield performance improvements for NER tasks. While rule-based filtering improved the F1 score by 0.5%, the adoption of more sophisticated post-processing strategies, such as probabilistic methods or ensemble techniques, holds the potential to further enhance the results.

Finally, the proposed systems secured the sixth and fifth ranks in the LegalLens NER and LegalLens NLI tasks, respectively, demonstrating their competitiveness in the shared task.

## 7 Acknowledgement

All experiments were carried out using Kaggle notebooks, which were equipped with a single NVIDIA P100 GPU. We extend our gratitude to Kaggle for their support in providing this computational resource.

## References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. Falcon-40b: an open large language model with state-of-the-art performance.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. Legallens: Leveraging llms for legal violation identification in unstructured text. *arXiv preprint arXiv:2402.04335*.

Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. 1999. An algorithm that learns what's in a name. *Machine learning*, 34:211–231.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Nyu: Description of the mene named entity system as used in muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6065–6075.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

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, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2021b. Generate, annotate, and learn: Generative models advance self-training and knowledge distillation. *arXiv:2106.06168*.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. A synthetic data approach for domain generalization of nli models. *arXiv preprint arXiv:2402.12368*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jingjing Huo, Kezun Zhang, Zhengyong Liu, Xuan Lin, Wenqiang Xu, Maozong Zheng, Zhaoguo Wang, and Song Li. 2023. AntContentTech at SemEval-2023 task 6: Domain-adaptive pretraining and auxiliary-task learning for understanding Indian legal texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 402–408, Toronto, Canada. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2211.03442*.

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.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. 2023. Semeval 2023 task 6: Legaleval-understanding legal texts. *arXiv preprint arXiv:2304.09548*.

Dhanachandra Ningthoujam, Pinal Patel, Rajkamal Kareddula, and Ramanand Vangipuram. 2023. Researchteam_hcn at semeval-2023 task 6: A knowledge enhanced transformers based legal nlp system. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1245–1253.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.(nips), 2017. *Advances in Neural Information Processing Systems*, 10:S0140525X16001837.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer. *Preprint*, arXiv:2311.08526.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition.

## A Example Appendix

**Prompts** Figure 2 showcases few-shot learning approach to generate NER data points using three randomly selected examples from the training dataset.

Figure 3 and 4 showcase prompts to generate praphrases of premise and hypothesis of the NLI training dataset.

```
<s>[INST]**Objective:**
Produce a realistic social media post about legal violations that include
clearly identified named entities. Each entity should be meticulously labeled
according to its type for straightforward extraction.

**Format Requirements:**
- The output should be formatted in JSON, containing the text and the
corresponding entities list.
- Each entity in the text should be accurately marked and annotated in the
'entities' list.
- Meticulously follow all the listed attributes.

**Entity Annotation Details:**
- All entity types must be in uppercase. For example, use "TYPE" not "type".
- Entity types must be one of the four types:
1. LAW: law or regulation breached
2. VIOLATION: content describing the violation
3. VIOLATED BY: entity committing the violation)
4. VIOLATED ON: victim or affected party

- There should not be multiple entities of each type. If there are multiple
entities of VIOLATION, it should be mentioned as one single span

**Output Schema: **

<start>
{
  "text": "{text content}",
  "entities": [
    {"entity": "entity name", "types": ["type"]},
    ...
  ]
}
<end>

**Here are some real-world examples**:

{examples}


[\INST]
```

Figure 2: Prompt design for NER dataset with task description and few-shot examples

```
Produce a realistic paraphrasing of the given court settlement. It should
retain all the details.  The output should strictly consist of the rephrasing
and nothing else.

**Here's the text**

{text}
```

Figure 3: Prompt design to paraphrase the premise of the NLI training dataset.

```
Produce a realistic paraphrasing of the given social media post. It should
retain all the details. The output should strictly consist of the rephrasing
and nothing else.

**Here's the text**

{text}
```

Figure 4: Prompt design to paraphrase the hypothesis of the NLI training dataset.