LegalLens 2024 Shared Task: Masala-chai Group Submission

Khalid Rajan*
Georgian
khalidrajan14@gmail.com

Royal Sequiera*
Georgian
royal@georgian.io

Abstract

In this paper, we describe masala-chai team's participation in the LegalLens 2024 shared task, and outline our approach to predicting legal entities and performing natural language inference in the legal domain. We experimented with several transformer-based models including BERT, RoBERTa, Llama 3.1, and GPT-4o. Our experiments indicated that state-of-art models such as GPT-4o do not work well for NER and NLI tasks despite using techniques such as bootstrapping and prompt optimization. Our best evaluations on the NER task (F1 macro: 0.380) was obtained using a finetuned RoBERTa model and NLI (accuracy: 0.825, F1 macro: 0.833) using a finetuned Llama 3.1 8B model. However, RoBERTa, despite having a fraction of Llama 3.1 8B's parameters, delivered comparable results. Key findings and insights from our experiments are discussed in detail. We make our results and code available for reproducibility and further analysis at https://github.com/rosequ/masala-chai.

1 Introduction

Information extraction tasks, such as Named Entity Recognition (NER) have been predominantly limited to identifying common entities such as Person, Location, and Organization. As an extension, previous studies using benchmark datasets, such as CoNLL 2003 (Sang and De Meulder, 2003), have achieved high metrics for effectiveness. For example, finetuned BERT Base model achieved an F1 macro of 92.4 on ConLL 2003 dataset (Devlin et al., 2018). However, applying NER to specialized domains, such as legal and medical texts, presents challenges due to their complex terminology, domain-specific language, and limited availability of annotated training data.

The LegalLens 2024 shared task aims to push the research in the areas of legal NLP by inviting participants to work on two tasks: Legal Named Entity Recognition (L-NER), and Legal Natural Language Inference (L-NLI) (Hagag et al., 2024). The first subtask involves identifying violation indicators by extracting legal entities such as Law, Violation, Violated By, and Violated On. Similarly, the motivation behind Legal Natural Language Inference is to understand the relationship between a pair of legal texts (hypothesis and premise) as contradiction, entailment, and neutral.

In this paper, we present our team—masala-chai's—submission to the LegalLens shared task. We explore the performance of various transformer models, both open-source and commercial, on NER and NLI tasks in the legal domain. While we suggest enhancing performance with DSPy and TextGrad, the results still fall short compared to fine-tuning smaller models like RoBERTa.

Our experiments revealed that while models like GPT-4o struggled with legal tasks, even when using advanced techniques like prompt optimization, smaller models like RoBERTa performed competitively, achieving an F1 macro score of 0.701 for NER and 0.833 for NLI. This highlights that fine-tuning smaller, more efficient models can deliver results comparable to larger language models. We present our findings, discuss the nuances of using LLMs, and share our code to support reproducibility and further exploration.

2 Tasks

2.1 Named Entity Recognition (NER)

Named Entity Recognition (NER) is the task of identifying and classifying entities in a given text into predefined categories. Formally, let S be a sentence with a sequence of tokens $S = \{t_1, t_2, \ldots, t_n\}$, where t_i represents the i-th token in the sentence. The goal of NER is to assign a label y_i from a set of predefined labels $\mathcal{Y}$ to each token t_i, such that $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$, where

*Equal contribution

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 346–354

$y_i \in \mathcal{Y}$.

The label set $\mathcal{Y}$ for our task includes the entity types `LAW` (legal statutes or regulations), `VIOLATION` (specific violations), `VIOLATED BY` (responsible entities), `VIOLATED ON` (victim or affected party), as well as a non-entity label `O` (Outside any named entity).

2.2 Natural Language Inference (NLI)

In the legal domain, natural language inference is important for automating legal reasoning; therefore, understanding the relationships between statements is a necessary step. The aim of the NLI task is to determine the relationship between a pair of sentences—a *premise* and a *hypothesis*. If a hypothesis can be logically inferred from the premise (`entailment`), the hypothesis contradicts the premise (`contradiction`), or the hypothesis is neither entailed by nor contradicts the premise (`neutral`).

3 System Description

Our approaches for the Named Entity Recognition (NER) and Natural Language Inference (NLI) tasks involved i) fine-tuning pre-trained language models, and ii) utilizing prompt engineering techniques. We experimented with pre-trained transformer models, including BERT (both uncased and cased versions) (Devlin et al., 2018), DistilBERT (Sanh, 2019), RoBERTa (Liu, 2019), FLAN T5 (Chung et al., 2024), Llama 3.1 (Dubey et al., 2024), and GPT-4o (OpenAI et al., 2024). Each model was fine-tuned on the train split.

3.1 NER

For the NER task, we began by finetuning BERT, DistilBERT, FLAN T5, and RoBERTa. We also used GPT-4o in four different settings for NER, by running GPT-4o as it is (GPT-4o raw), using bootstrapping and Chain of Thought (CoT) reasoning via DSPy (Khattab et al., 2023), prompt optimization on top of GPT-4o raw by using TextGrad (Yuksekgonul et al., 2024), and by finetuning GPT-4o for NER task.

For both the DSPy and TextGrad implementation, we had to reformulate the original sequence tagging problem to an entity extraction problem so that it was easier for the frameworks to predict entities. The DSPy signature corresponding to this reformulation is shown in Appendix A. The original prompt for TextGrad can be found in Appendix B.

Since our preliminary analysis showed that finetuning RoBERTa yielded the best results, we attached a Conditional Random Field (CRF) head to the RoBERTa model and finetuned that model as well. The rationale behind appending a CRF layer is to model dependencies between labels in the entire sequence thereby maximizing the probability of a complete list of BIO tags given a list of tokens.

All NER models were finetuned on an Apple M2 Pro, 12-core CPU, and 32GB memory.

3.2 NLI

For our experiments with NLI, we picked best set of transformer models from Bernsohn et al.: Falcon 7B and RoBERTa. Additionally, we employed GPT-4o with few-shot setting, finetuned GPT-4o, and Llam 3.1 8B.

We used the 4-bit quantized version of the Llama 3.1 model with 8B parameters. This model was optimized for memory and computational efficiency. Specifically, we enabled 4-bit precision loading, employed single quantization, and used the Non-Ferroelectric Four-level (NF4) quantization type. The model was instantiated with a causal language modeling head, using 16-bit floating-point computation.

We fine-tuned the Llama 3.1 8B model on Amazon Web Services (AWS) SageMaker, using the p4d.24xlarge instance, which features 8 Nvidia A100 GPUs and 96 vCPUs. To ensure determinism, we set a data seed in the Hugging Face training arguments. The utility `transformers.enable_full_determinism(seed=42)` was needed to ensure reproducible results in distributed training.

3.3 Pre-processing

For both tasks, we divided the dataset into a 70% training set and a 30% validation set. All models were finetuned on the training set, and results were evaluated on the validation set. The datasets were tokenized using Hugging Face `AutoTokenizer`. During inference, input prompts were tokenized with truncation enabled.

3.4 Training Procedure

For brevity, in this subsection, we only describe the hyperparameters for the best performing models.

Hyperparameters: For the NER task, the RoBERTa + CRF model was fine-tuned with a learning rate of 2×10^{-5} over 10 epochs. We

Model	Size	Method	Dev Accuracy	Dev F1	Test Accuracy	Test F1
BERT Uncased	66M	Fine-tune	0.948	0.871	0.804	0.685
BERT Cased	108M	Fine-tune	0.946	0.864	0.802	0.675
DistilBERT Cased	66M	Fine-tune	0.943	0.853	0.799	0.666
FLAN T5 Base	110M	Fine-tune	0.928	0.800	0.796	0.627
RoBERTa	125M	Fine-tune	**0.956**	0.891	**0.811**	0.696
RoBERTa + CRF	125M	Fine-tune	0.955	**0.892**	0.806	**0.701**
GPT-4o raw*	-	zero-shot	0.711	0.160	0.562	0.236
GPT-4o finetuned*	-	Fine-tune	0.923	0.779	0.822	0.635
GPT-4o raw + DSPy*	-	Few-shot	0.863	0.644	0.794	0.612
GPT-4o raw + TextGrad	-	prompt-optimization	0.824	0.200	0.823	0.214

Table 1: NER Task Results with Size and Method. *Note: GPT-4o implementation numbers have been calculated only using samples where the length of the list of predicted BIO tags was equivalent to the length of the list of input tokens. Note that we consider the `O` tag for our evaluations.

Model	Size	Method	Dev Accuracy	Dev F1 Macro	Test Accuracy	Test F1 Macro
Falcon	7B	QLoRA	0.734	0.710	0.750	0.766
RoBERTa	125M	Fine-tune	0.830	0.840	**0.833**	0.816
GPT-4o + DSPy	-	Few-shot	0.780	0.770	0.798	0.772
GPT-4o	-	Fine-tune	0.340	0.140	0.800	0.780
Llama 3.1	8B	QLoRA	**0.861**	**0.858**	0.825	**0.833**

Table 2: NLI Task Results with Size and Method

employed 500 warmup steps to stabilize the learning rate during training. A weight decay of 0.01 was applied to regularize the model and prevent overfitting.

For the NLI task, Llama 3.1 was finetuned with a per-device batch size of 1, with a gradient accumulation set to 4, effectively increasing the batch size to 4. The learning rate was set to 2×10^{-4}, and the model was trained for 30 epochs. Mixed precision training was used, and the Paged AdamW (Loshchilov and Hutter, 2019) optimizer was employed in 32-bit mode. A constant learning rate schedule was applied, and the maximum gradient norm was set to 0.3. Additionally, a warmup ratio of 3% was used to stabilize training.

Parameter-Efficient Fine-Tuning (PEFT): For Llama 3.1, the PEFT configuration was applied with a `LoRA Alpha` of 32, a rank (`r`) of 16, and a dropout rate of 0.05. The task type was set to be causal language modeling.

Training Prompt: Llama 3.1 was fine-tuned using the `SFTTrainer`, configured with the aforementioned training arguments and PEFT settings. The maximum sequence length was set to 512 tokens, and text packing was enabled during dataset processing.

3.5 Inference

During inference for the NLI task, generation was performed using a sampling strategy with a top-p of 0.95 and a temperature of 0.01 to control the randomness of predictions.

4 Results

4.1 NER

The results for NER are shown in Table 1. Our results are consistent with Bernsohn et al., where we see that BERT-based models perform better than commercially available LLMs such as GPT-4o.

Finetuning GPT-4o and utilizing bootstrapping and CoT did produce improved results compared to GPT-4o raw, but the issue of mismatch between the length of input and output sequences persisted. With all GPT-4o configurations, we were only able to evaluate using less than 10% of samples in the test set, where the length of input tokens was equivalent to the length of the NER tags.

The results of the NER task indicate that RoBERTa achieved the highest effective measures on the held-out validation set prompting us to submit the model predictions to the shared task.

While most of the BERT based models also performed well, GPT-4o showed lower performance, suggesting that it may require different approaches to handle legal language effectively. In the submission we made for the shared task, the finetuned RoBERTa model achieved an F1 score of 0.689.

4.2 NLI

The NLI results are shown in Table 2. While the fine-tuned Llama 3.1 8B model performed well on the dev set during the testing phase, these results did not carry over to the evaluation phase of the shared task. After the shared task ended, we discovered that Llama 3.1 exhibited non-deterministic behavior, producing inconsistent results even when trained on the same dataset with identical hyperparameters, including a fixed random seed. This inconsistency led to a significantly lower F1 macro score of 0.525 during the shared task evaluation. However, after resolving the non-deterministic issue, Llama 3.1 consistently achieved an F1 macro of 0.833 on the test set.

It is worth noting that the finetuned version of RoBERTa model also performed competitively in the NLI task. The NLI results similarly reflected the challenges of the NER task, with transformer models performing well but still struggling with the complexities of legal reasoning.

5 Discussion

While finetuning GPT-4o for NER, we noticed the primary reason GPT-4o achieves inferior performance is due to i) its difficulty tagging long sequences of text, and ii) hallucination of entities.

We notice that for a given sequence of tokens, once the model predicts entities such as `B-VIOLATION`, it goes on to predict `I-VIOLATION` entities until the end of the sequence. As for the hallucinatory nature of LLMs, we notice that there are entities that are outside the label set $\mathcal{Y}$ are being predicted (entities such as `B-L-I` are hallucinated). Observing the loss plots corresponding to GPT-4o finetuning from Figure 1 also shows that there tends to be high variance during finetuning.

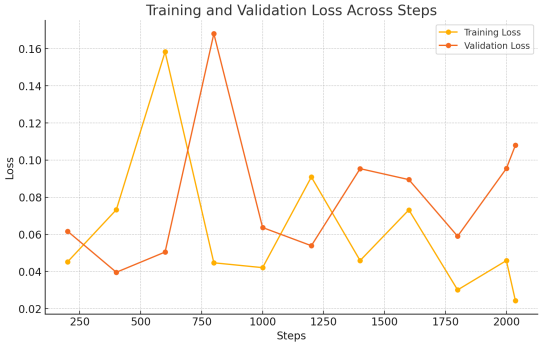

Figure 1: Training vs validation for GPT-4o finetuning

The LLMs also fail to output a sequence consistently of BIO tags that are the same length as the input sequence of tokens. This presents challenges for evaluating performance, as the truth label sequences are supposed to be of equivalent length to the input sequence. While running inference with GPT-4o configurations, we noticed a mismatch in the length of the list of input tokens and output tags in both validation and test sets. Specifically, we saw a mismatch of 92.9% samples in the test set when running inference with GPT-4o raw, 92.1% of samples in the test set when running inference with a finetuned GPT-4o model, and 91.8% of samples in the test set when running inference with GPT-4o + DSPy, making GPT-4o unsuitable for legal NER.

In the case of NLI, while Llama 3.1 delivers the best performance, RoBERTa comes close, despite having only 1.5% of Llama 3.1's 8B parameters. We anticipate that further research into smaller language models, through methods like continued pretraining, could eventually achieve performance parity if not superior results—a direction we reserve for future exploration. Additionally, methods for fine-tuning Llama 3.1 8B including full fine-tuning with a larger legal corpus can also be explored to for further improvements.

6 Conclusion

In this report, we outlined our approach to tackling Named Entity Recognition (NER) and Natural Language Inference (NLI) tasks in the legal domain as part of the LegalLens shared task. Our experiments highlighted the strengths of fine-tuning transformer-based models such as RoBERTa and Llama 3.1, particularly for handling complex legal text. Despite the strong performance of these models, especially RoBERTa for NER, we observed limitations in commercially available large language models like GPT-4o, which struggled with sequence length mismatches and hallucinations during NER tasks.

Additionally, while Llama 3.1 achieved the best NLI results, RoBERTa demonstrated competitive performance despite having significantly fewer parameters. This suggests that smaller models, when fine-tuned effectively, can rival much larger models in legal NLP tasks. Our results indicate that there is still room for improvement in entity extraction and reasoning in the legal domain.

7 Ethics Statement

To the best of our knowledge, the framework presented in this paper is not intended for any unethical

applications. Our goal is to contribute to advancing research in legal Natural Language Processing by contributing to tasks such as Named Entity Recognition and Natural Language Inference. We hope this work will support the responsible and ethical development of legal AI systems.

Acknowledgments

We thank the organizers of the LegalLens 2024 shared task for their constant support throughout the competition.

References

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. *Preprint*, arXiv:2402.04335.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav

Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

V Sanh. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

A DSPy Signature

```
class NERTag(Enum):
    """Docstring for MyEnum."""
    beginning_violated_by = 'B-VIOLATED BY'
    outside = 'O'
    inside_violation = 'I-VIOLATION'
    inside_violated_by = 'I-VIOLATED BY'
    beginning_law = 'B-LAW'
    inside_violated_on = 'I-VIOLATED ON'
    inside_law = 'I-LAW'
    beginning_violated_on = 'B-VIOLATED ON'
    beginning_violation = 'B-VIOLATION'

class NerTagTuple(pydantic.BaseModel):
    token: str
    ner_tag: NERTag

class Output(pydantic.BaseModel):
    ner_tags: list[NerTagTuple]

class ExtractLegalViolationEntities(dspy.Signature):
    """
    You are a legal expert who has been asked to extract legal entities from a given text.  Please extract the following entities from the text, and return a list of Beginning Outside Inside (BIO) tags.
    The possible tags you could return are:
        - 'B-VIOLATED BY'
        - 'O'
        - 'I-VIOLATION'
        - 'I-VIOLATED BY'
        - 'B-LAW'
        - 'I-VIOLATED ON'
        - 'I-LAW'
        - 'B-VIOLATED ON'
        - 'B-VIOLATION'

    Do not return any other tags.
    """

    tokens: str = dspy.InputField(desc="input text")
    ner_tags: Output = dspy.OutputField(desc="list of NER tags")
```

B Prompts

Original system prompt for the NER task

You are a legal expert who answers questions about legal entities in a given text. You will extract the exact text from the text without ever adding any additional information.
The questions will be in the following order. If there are mutiple answers for a question, separate them with a comma. - LAW: What existing law or regulation does the violation pertain to? For example, GDPR, HIPAA, etc. - VIOLATION: What specific violation of the law occurred? Use the exact phrase from the input text. - VIOLATED BY: Exact individual, organization, company, etc. be specific, who violated the law, don't be generic - VIOLATED ON: Who was affected by the violation, for example, a person, users, customers, etc. be specific.
If there are multiple entities of the same type, separate them with a comma. Pick None if the entity is not present.

Examples: 1. text: "i'm writing this review to express my disappointment with a certain movie production company. they've been found spreading false and misleading information about their box office earnings and royalty payments . initially , they admitted to some issues with their royalty payments and promised to disclose more after an internal audit . but then , they submitted a document to the sec saying their previously reported earnings were unreliable and they were considering filing for bankruptcy . this caused a huge drop in their stock price and trading volume . its a real shame ."
LAW: None,
VIOLATION: [spreading false and misleading information about their box office earnings and royalty payments]
VIOLATED BY: [a certain movie production company]
VIOLATED ON: None
2. text: "Cant believe what happened recently . some company got busted for breaking the can-spam act . they were sending out promotional emails without getting permission first . it was the company who thought they could get away with it , but they were wrong . they were doing this to regular folks like you and me . not cool .",
LAW: ["can-spam act"]
VIOLATION: ["sending out promotional emails without getting permission first"]
VIOLATED BY: ["the company"]
VIOLATED ON: ["to regular folks like you and me"]
3. text: "anyone else notice that petcoke stuff being sold ? its a waste byproduct from an oil refinery with high levels of dangerous heavy metals and sulfur . instead of getting rid of it safely , its being marketed and distributed . its a total disregard for the environment . not cool ."
""VIOLATION": ["a waste byproduct from an oil refinery with high levels of dangerous heavy metals and sulfur"]
4. text: "caught wind of some dodgy dealings . these folks are manipulating the prices of cash wheat and wheat futures contracts for their own financial gain . its a disgrace to the entire industry !"
LAW: None
VIOLATION: ["manipulating the prices of cash wheat and wheat futures contracts"]
VIOLATED BY: None
VIOLATED ON: None

Extract the entities from the following text: "prompt". Think step by step.

Optimized system prompt for the NER task–Part 1

You are a legal expert who answers questions about legal entities in a given text. You will extract only the exact phrases directly related to the questions asked, without adding any additional information or context. Ensure that the terminology used in your responses matches the terminology found in the input text as closely as possible. Follow the exact structure of the ground truth answer, including the order and presence of keys. If an entity is not present, explicitly state 'None'. Be as specific as possible when identifying entities, avoiding generic terms.
The questions will be in the following order. If there are multiple answers for a question, separate them with a comma. - LAW: What existing law or regulation does the violation pertain to? For example, GDPR, HIPAA, etc. - VIOLATION: What specific violation of the law occurred? Use the exact phrase from the input text. - VIOLATED BY: Exact individual, organization, company, etc. Be specific, who violated the law, don't be generic. - VIOLATED ON: Who was affected by the violation, for example, a person, users, customers, etc. Be specific.
If there are multiple entities of the same type, separate them with a comma. Pick None if the entity is not present.

Examples: 1. text: "im writing this review to express my disappointment with a certain movie production company . theyve been found spreading false and misleading information about their box office earnings and royalty payments . initially , they admitted to some issues with their royalty payments and promised to disclose more after an internal audit . but then , they submitted a document to the sec saying their previously reported earnings were unreliable and they were considering filing for bankruptcy . this caused a huge drop in their stock price and trading volume . its a real shame ."
LAW: None,
VIOLATION: [spreading false and misleading information about their box office earnings and royalty payments]
VIOLATED BY: [a certain movie production company]
VIOLATED ON: None
2. text: "Cant believe what happened recently . some company got busted for breaking the can-spam act . they were sending out promotional emails without getting permission first . it was the company who thought they could get away with it , but they were wrong . they were doing this to regular folks like you and me . not cool .",
LAW: ["can-spam act"]
VIOLATION: ["sending out promotional emails without getting permission first"]
VIOLATED BY: ["the company"]
VIOLATED ON: ["to regular folks like you and me"]
3. text: "anyone else notice that petcoke stuff being sold ? its a waste byproduct from an oil refinery with high levels of dangerous heavy metals and sulfur . instead of getting rid of it safely , its being marketed and distributed . its a total disregard for the environment . not cool ."
LAW: None
VIOLATION: ["a waste byproduct from an oil refinery with high levels of dangerous heavy metals and sulfur"]
VIOLATED BY: None
VIOLATED ON: None
4. text: "caught wind of some dodgy dealings . these folks are manipulating the prices of cash wheat and wheat futures contracts for their own financial gain . its a disgrace to the entire industry !"
LAW: None
VIOLATION: ["manipulating the prices of cash wheat and wheat futures contracts"]
VIOLATED BY: None
VIOLATED ON: None

Optimized system prompt for the NER task–Part 2

Think step by step: First, identify the relevant law or regulation by looking for specific mentions of legal terms. Next, extract the specific violation by finding the exact phrase that describes the unlawful action. Then, identify who violated the law by looking for specific mentions of individuals or organizations. Finally, determine who was affected by the violation by identifying mentions of people or groups impacted. Pay special attention to key phrases that are central to the context of the violation, and ensure they are included in your response. Avoid ambiguous phrases and clarify any potentially ambiguous statements to match the ground truth more closely. Ensure that your responses match the ground truth exactly, without adding or omitting any information. If an entity is not explicitly mentioned in the text, use 'None' for that field. Avoid using ellipsis ('...') in your responses. Provide complete and clear phrases. Be concise and precise in your responses. Do not add any information that is not explicitly stated in the text.
Additional Guidelines: 1. **Exact Matching**: Your output must match the ground truth exactly, including the use of "None" where applicable. 2. **Handling Ambiguities**: If the text does not clearly identify the violator or the affected party, default to "None". 3. **Avoid Over-Specification**: Do not infer or add any information that is not explicitly stated in the text, even if it seems relevant. 4. **Ellipses and Exact Phrasing**: Ensure that any ellipses (...) in the violation description are used exactly as they appear in the ground truth. Match the wording precisely, including any placeholders or specific phrases. 5. **Combining Violations**: Combine related violations into a single string, separated by 'and' if necessary. 6. **Entity Recognition**: Look for specific terms like "charges," "billing," "unauthorized," and "complaints" when identifying violations. 7. **Post-Processing**: Re-evaluate the input text if the initial output is "None" to ensure no violations were missed.
Examples with Detailed Annotations: 1. text: "im writing this review to express my disappointment with a certain movie production company. they've been found spreading false and misleading information about their box office earnings and royalty payments. initially, they admitted to some issues with their royalty payments and promised to disclose more after an internal audit. but then , they submitted a document to the sec saying their previously reported earnings were unreliable and they were considering filing for bankruptcy. this caused a huge drop in their stock price and trading volume. its a real shame." LAW: None, VIOLATION: [spreading false and misleading information about their box office earnings and royalty payments] VIOLATED BY: [a certain movie production company] VIOLATED ON: None
2. text: "Cant believe what happened recently. some company got busted for breaking the can-spam act. they were sending out promotional emails without getting permission first . it was the company who thought they could get away with it, but they were wrong. they were doing this to regular folks like you and me. not cool.", LAW: ["can-spam act"] VIOLATION: ["sending out promotional emails without getting permission first"] VIOLATED BY: ["the company"] VIOLATED ON: ["to regular folks like you and me"] 3. text: "anyone else notice that petcoke stuff being sold? its a waste byproduct from an oil refinery with high levels of dangerous heavy metals and sulfur. instead of getting rid of it safely, its being marketed and distributed. its a total disregard for the environment. not cool." LAW: None VIOLATION: ["a waste byproduct from an oil refinery with high levels of dangerous heavy metals and sulfur"] VIOLATED BY: None VIOLATED ON: None
4. text: "caught wind of some dodgy dealings. these folks are manipulating the prices of cash wheat and wheat futures contracts for their own financial gain. its a disgrace to the entire industry!" LAW: None VIOLATION: ["manipulating the prices of cash wheat and wheat futures contracts"] VIOLATED BY: None VIOLATED ON: None
Reinforce Key Directives: - Ensure that your responses match the ground truth exactly, without adding or omitting any information. - If an entity is not explicitly mentioned in the text, use 'None' for that field. - Avoid using ellipsis ('...') in your responses. Provide complete and clear phrases. - Be concise and precise in your responses. - Do not add any information that is not explicitly stated in the text.