# Semantists at LegalLens-2024: Data-efficient Training of LLM's for Legal Violation Identification

**Rajaraman Kanagasabai**
Institute for Infocomm Research,
Agency for Science, Technology and Research
1 Fusionopolis Way
Singapore 138632,
kanagasa@i2r.a-star.edu.sg

**Hariram Veeramani**
University of California,
Los Angeles,
USA.
hariram@ucla.edu

## Abstract

In this paper, we describe our system for LegalLens-2024 Shared Task on automatically identifying legal violations from unstructured text sources. We participate in Subtask B, called Legal Natural Language Inference (L-NLI), that aims to predict the relationship between a given premise summarizing a class action complaint and a hypothesis from an online media text, indicating any association between the review and the complaint. This task is challenging as it provides only limited labelled data. In our work, we adopt LLM based methods and explore various data-efficient learning approaches for maximizing performance. In the end, our best model employed an ensemble of LLM's fine-tuned on the task-specific data, and achieved a Macro F1 score of 78.5% on test data, and ranked 2nd among all teams submissions.

## 1 Introduction

Legal violation identification is an important problem that aims to automatically uncover legal violations from unstructured text sources and assign potential victims to these violations. In the past, several works have addressed this but often relied on specialized models tailored for specific domain applications (Chalkidis et al., 2020; Yang, 2022). These models, while effective in their specific domains, lack the versatility needed to address the wide array of legal violations that can occur across different contexts.

The LegalLens Shared Task (Hagag et al., 2024) proposes to address the legal violation identification task using named entity recognition (NER), and the other for associating these violations with potentially affected individuals using natural language inference (NLI). In this paper, we report our system for addressing the NLI task (Sub-Task B).

Broadly, our approach is to adopt LLM based methods and explore various data-efficient learning approaches for maximizing performance on the NLI task. Our best model employed an ensemble of LLM's fine-tuned on the task-specific data. Our final system achieved a Macro F1 score of 78.5 and ranked 2nd among all teams submissions. Surprisingly, the classical Falcon LLM's outperformed many other SOTA LLM's. Also, our benchmark results highlight the challenge of these tasks and indicate there is ample room for model improvement. We demonstrate the limitation of the general LLM based methods and discuss possible future work.

## 2 Task and Dataset Description

The LegalLens challenge (Hagag et al., 2024) proposes two shared sub-tasks:

- Sub-Task A. Legal Named Entity Recognition (L-NER)
- Sub-Task B. Legal Natural Language Inference (L-NLI)

Participants can choose either of the two sub-task or both. We participate in Sub-Task B, defined as below.

### 2.1 Subtask B

*Legal Natural Language Inference (L-NLI)* Given a premise summarizing a class action complaint and a hypothesis from an online media text, the task is to determine if the relationship is entailed, contradicted, or neutral, indicating any association between the review and the complaint.

In contrast to NER which can help in detecting legal violations within unstructured textual data, the NLI task assists in associating these violations with potentially affected individuals.

### 2.2 Dataset

To facilitate the L-NLI task, the participants are provided with a dateset constructed based on previous class action cases and legal news. The latter is done by first summarizing the news to create the

*Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 355–360
November 16, 2024 ©2024 Association for Computational Linguistics

| Domain | Labels<br>#E/#C/#N | #Samples |
|---|---:|---:|
| Consumer Protection | 16/17/29 | 62 |
| Privacy | 56/54/53 | 163 |
| TCPA | 26/27/21 | 74 |
| Wage | 6/3/4 | 13 |
| Total | 104/101/107 | 312 |

Table 1: Distribution of L-NLI Task Training Data, including the number of samples (column 3) and the class distribution (column 2) under each legal domain, where the classes ’Entailed’ (E), ’Contradicted’ (C), and ’Neutral’ (N) are denoted using their first letters respectively.

premise, and generating a hypothesis using GPT-4 (Achiam et al., 2023) and subsequently validated by domain experts.

The data covers 4 legal domains namely Consumer Protection, Privacy, TCPA and Wage. In total, the data comprises 312 labeled samples (See Table 1). This is clearly small in size which makes the task quite challenging due to the risk of overfitting and limited generalization.

## 3 Our Approach

Natural language inference (NLI) is the task of detecting inferential relationships between a premise text and a hypothesis text (Dagan et al., 2010; Romanov and Shivade, 2018; Storks et al., 2019), which is considered fundamental in natural language understanding (NLU) research (Bowman et al., 2015). In L-NLI task, the premise is a summary of a class action complaint and the hypothesis an online media text, and the objective is to determine if the relationship is ’entailed’, ’contradicted’, or ’neutral’.

Several NLI systems have been proposed in the literature (Bowman et al., 2015; Storks et al., 2019), and can be adapted for the L-NLI task. (Bernsohn et al., 2024) investigated this by fine-tuning popular Small language models, such as BERT and RoBERTa, and reported that the models struggled with the task. Also, using their legal counterparts, like Legal-BERT (Chalkidis et al., 2020), Legal- RoBERTa (Chalkidis et al., 2023), and Legal- English-RoBERTa (Niklaus et al., 2023) models also did not lead to much improvements. This can be attributed to the small data, as most of the models typically assume sufficiently large number of labelled data. This is particularly true for NLI which is essentially a 3-way sentence pair classification problem.

In comparison, LLMs are reported to learn relatively better in low data situations and generalize well to out-of-distribution (OOD) test data sets (Brown et al., 2020). This is in part due to their pre-training on variety of datasets, eg. SNLI and MNLI, as supported by the preliminary results of (Bernsohn et al., 2024) using fine-tuned Falcon (Almazrouei et al., 2023) and Llama (Touvron et al., 2023) models.

In our work, we perform a more extensive study by considering more LLM’s and explore various LLM based strategies and techniques, beyond prompt engineering, for maximizing performance in the given task.

### 3.1 Vanilla Fine-tuning of LLM’s

We consider several popular LLM’s and fine-tune the models using the task-specific labeled data. This helps in adjusting the parameters of a pre-trained large language models to the L-NLI task. However, as the training data is too sparse, we do not use full fine-tuning but instead resort to Parameter Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019). PEFT is a technique used to improve the performance of pre-trained LLMs on specific downstream tasks while minimizing the number of trainable parameters. It offers a more efficient approach by updating only a minor fraction of the model parameters during fine-tuning. PEFT technique selectively modifies only a small subset of the LLM’s parameters, typically by adding new layers or modifying existing ones in a task-specific manner. This approach significantly reduces the computational and storage requirements while maintaining comparable performance to full fine-tuning. We adopt QLoRA (Dettmers et al., 2024), which applies a low-rank approximation to the weight update matrix and also quantizes the weights of the LoRA (Hu et al., 2021) adapters resulting in reduced memory footprint and storage requirements.

### 3.2 Data Augmentation

Data augmentation involves the adoption of new methods aimed at improving model efficacy by enriching training data diversity without necessitating further data collection efforts. Data augmentation using LLMs has heralded innovative learning paradigms, marking a significant departure from traditional methods (Ding et al., 2024). In our case, we employ data augmentation to address the challenging categories such as 'Wage' which has only 13 samples in total. We explore various strategies including 1. Data creation which leverages the few-shot learning ability of LLMs to create a large synthetic dataset; 2. Data labeling which uses the LLM to label existing datasets; 3. Data reformation which transforms existing data to produce new data.

### 3.3 LLM Ensembleing

Open-source LLMs exhibit diverse strengths and weaknesses due to variations in data, architectures and hyperparameters, making them complementary. Often there does not exist one LLM that dominates the competition for all examples. Thus, it is attractive to ensemble the output of the best LLMs (based on input, task and domain) to give consistently superior performance across examples. By combining their unique contributions; the biases, errors and uncertainties in individual LLMs can be alleviated, resulting in outputs aligned with human preferences (Rajamanickam and Rajaraman, 2023; Yang et al., 2023).

In our experiments, we pool multiple LLM's and explore ensembling their individual predictions using voting strategies to improve overall robustness and accuracy.

Our experiments and results are described in detail in the following section.

## 4 Experiments & Results

In the L-NLI challenge dataset, there are 312 instances distributed across 4 legal domains (See Table 1). For training the models, we randomly split the instances into 80% training and 20% validation, repeat the experiments five times and report the average performance. The results are presented below, where performance metrics are quoted in % for easier interpretation, unless stated otherwise.

The experiments were executed on NVIDIA-GeForce RTX 2080 series with eight cores of GPU machines with 8*12 GB of memory for all our experiments. Also, to train T5 large models, we have used NVIDIA-GeForce Tesla V100 series SXM2-32GB with 5 cores of GPU machines. Models were trained for 3-5 hours for training and reasoning. The pretrained weights for the transformers prior to fine-tuning were from the HuggingFace NLP Library.

| Model | Val (Macro F1) |
|---|---|
| T5-base | 81.4 |
| Falcon-7b | 88.4 |
| Llama2-7b | 86.6 |
| Gemma-7b | 84.5 |
| Data Augmentation | 84.5 |
| LLM Ensembling | **89.4** |

Table 2: Comparison of performance of the models explored in our experiments.

### 4.1 Approach 1: (Vanilla Fine-tuning of LLM's) :

We first evaluated a basic fine-tuning approach. We considered SOTA LLM's such as Llama2-7b (Touvron et al., 2023), Gemma-7b (Team et al., 2024). Additionally, we included Falcon-7b (Almazrouei et al., 2023) as it had good reported performance (Bernsohn et al., 2024). Also included was T5 (Raffel et al., 2020) to serve as a baseline.

We adopted QLoRA fine-tuning, which applies a low-rank approximation to the weight update matrix and also quantizes the weights (Dettmers et al., 2024). For parameter settings, we use a QLoRA rank of 64, alpha of 32, and trained the models for 20 epochs with an initial learning rate of 2e-4, and a dropout rate of 0.25.

We observed that Falcon-7b performed surprisingly better than SOTA models like Llama and Gemma, and achieved 88.4% on validation data. Hence we decided to adopt it for further studies.

### 4.2 Approach 2: (Data Augmentation) :

We considered Falcon-7b, and employed data augmentation specifically to address the challenging category 'Wage' which has only 13 samples in total. In particular, we adopted data creation, labeling and reformation strategies to augment the training data, as below.

Using GPT-4 (Achiam et al., 2023), we created additional data using prompt engineering by first leveraging the few-shot learning ability to create synthetic samples and labels. Then, for each source sample, say 'Entailed', transform existing sample to produce samples for 'Contradict' and 'Neutral'. Thus we triple the labeled data and use random sampling to create train/val sets.

However, fine-tuning with the augmented data resulted in F1 score below that of unaugmented data, and so we abandoned it.

| Domain | Labels<br>#E/#C/#N | #Samples |
|---|---|---|
| BIPA | 14/4/4 | 22 |
| Consumer | 4/1/3 | 8 |
| Data Breach | 8/5/7 | 20 |
| TCPA | 5/2/2 | 9 |
| VPPA | 2/2/2 | 6 |
| Wage | 7/1/11 | 19 |
| Total | 40/15/29 | 84 |

Table 3: Distribution of L-NLI Test Data, including the number of samples (column 3) and the class distribution (column 2) under each legal domain, where the classes 'Entailed' (E), 'Contradicted' (C), and 'Neutral' (N) are denoted using their first letters respectively.

| Domain | #Correctly<br>Classified | #Misclassified |
|---|---|---|
| BIPA | 13 | 9 |
| Consumer | 7 | 1 |
| Data Breach | 18 | 2 |
| TCPA | 8 | 1 |
| VPPA | 4 | 2 |
| Wage | 13 | 6 |
| Total | 63 | 21 |

Table 4: Performance of our final model on the Test Data, across the 6 domains included in the data

### 4.3 Approach 3: (LLM Ensembling) :

Ensembling aims to combine the outputs of multiple LLMs (based on input, task and domain) so as to achieve better accuracy and robustness across all samples. Towards this, we trained 3 instances of Falcon-7b, each with a different set of randomly split (80-20) training data. (Ideally a partitioned data is preferred but due to small size we decided against it. ) We ran inference individually on the 3 models, and aggregated the predictions using majority voting. This ensemble approach achieved the best score (See Table 2), making it as our final submission.

We planned to perform extensive ensembling experiments using different LLM's, data sizes, etc. but could not complete them due to resource limitations. This deserves further study.

### 4.4 Analysis of Test Results

The test results and the data with target labels were announced soon after submission deadline. Our system achieved a Macro F1 score of 78.5% on test data, and ranked 2nd among all teams submissions.

Table 3 provides details about test data statistics. We note that the test data is from 6 domains, compared to 4 domains in training data. This clearly requires OOD performance, and possibly the reason for the significant drop in F1 from validation score.

As further investigation, we performed error analysis using two types of classification errors (Bernsohn et al., 2024): first-class errors, which involve confusions between "Contradict" and "Entailed", and second-class errors, which are misclassifications of "Contradict" or "Entailed" as "Neutral". Our final model had 21 Class-2 errors, and no Class-1 errors, which implies that the model has difficulty in identifying edge cases whether there is violation or not.

We present a distribution of errors across the domains in Table 4. The model performed well on Consumer, Data Breach and TCPA which had similar ones in training set. In contrast, the proportion of errors in the unseen domains BIPA and VPPA were significantly larger. Similar performance degradation was also observed for 'Wage' which can be recalled as one that had too few training samples.

In summary, we conclude that our LLM ensemble model performed fairly well for identifying legal violations, though there is scope for further improvements in tackling small data and OOD situations.

## 5 Discussion and Conclusion

This paper described our system for LegalLens-2024 Shared Task that aims to automatically uncover legal violations from unstructured text sources and assign potential victims to these violations. We participate in Subtask B, called Legal Natural Language Inference (L-NLI), that aims to predict the relationship between a given premise summarizing a class action complaint and a hypothesis from an online media text, indicating any association between the review and the complaint.

This task is challenging in view of the limited labelled data, and hence we explored various approaches for data-efficient learning with LLM's, such as PEFT fine-tuning, Data Augmentation and LLM Ensembling. In the end, our ensemble approach performed the best and achieved a Macro F1 score of 78.5%, and ranked 2nd among all teams submissions. The key findings are:

- LLM Fine-tuning improves zero-shot and few-shot performance. This possibly implies that specific domains can benefit from task specific training data even if smallish in size.

- The performance of various LLM's overall are somewhat close. Though Falcon emerged as the winner, the margins were not huge, and our T5 baseline was not far behind.

- Simple data augmentation may not be enough to guarantee improved performance. More careful data generation and possibly some human involvement is required.

- Ensemble approach has strong promise to achieve robust performance across all examples.

In summary, our research highlight the challenge of legal violation identification in real-life, and the limitations of SOTA LLM's. This further suggests that there is ample room for model improvement and scope for possible future work, especially under limited data settings.

## Limitations

Our work explored various LLM strategies for identifying legal violations under small data settings, but is clearly preliminary. We were limited by resource constraints and so could not do explore fine-tuning very large models (11b or bigger) or try other data augmentation experiments, along with extensive hyperparameter optimization. A more rigorous experimentation may be required to further validate the findings of the paper.

## Acknowledgments

This work was supported by fundings from A*STAR Epidemic Preparedness Horizontal Coordinating Office (EP HTCO) FY22_CF_HTCO SEED_EP_IDLabs_C22714020

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

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*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

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*.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using large language models: Data perspectives, learning paradigms and challenges.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

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.

Saravanan Rajamanickam and Kanagasabai Rajaraman. 2023. I2r at semeval-2023 task 7: Explanations-driven ensemble approach for natural language inference over clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1630–1635.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. 2023. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*.

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.