Lost in Legalese: NLP for Privacy Risk Detection

Stanford CS224N Custom Project

Ray Hu
Department of Computer Science
Stanford University
rayhu@stanford.edu

Benjamin Ward
Institute for Computational and Mathematical Engineering
Stanford University
wardb@stanford.edu

Basant Khalil
Department of Computer Science
Stanford University
bkhalil@stanford.edu

Abstract

Objective This project utilizes NLP techniques, specifically fine-tuned Legal-BERT, to assess privacy risks in Terms and Conditions (T&C) texts. By classifying individual clauses and assigning fairness ratings to entire documents, we aim to help users understand the implications of these texts on their privacy.

Motivation: Legal language is one of the most complex forms of natural language. A tragic 2024 case brought this issue to public attention: a woman passed away at Disney Park, and the company argued that her family had waived their right to sue because she had accepted the T&C when she signed up for a Disney+ trial back in 2019. The legal clause, buried in fine print, favored the company, highlighting how consumers often unknowingly forfeit their rights. Our project focuses on the analysis on the privacy risks in often deliberately obfuscated T&C texts.

Approach We fine-tune Legal-BERT using Low Rank Adaptation (LoRA) for individual clause classification (***Very Bad, Bad, Neutral, Good***) and document-level classifcation **fairness ratings (A–E)**. We compare these methods to our baseline of zero-shot prompting, and with further prompt engineering techniques.

Main Findings Fine-tuned **Legal-BERT** achieved **84% accuracy**, significantly outperforming **Llama 2 zero-shot (28.3%)**. Although finetuned **Legal-BERT (LoRA)** yielded less good results, LoRA fine-tuning enhanced fairness rating stability and explainability compared to general-purpose models. While prompt engineering improved Llama 2's performance, it remained notably weaker than domain-specific fine-tuning. However, ambiguous clauses remain a challenge, suggesting the need for additional legal context.

Future Work Integrating Retrieval-Augmented Generation (RAG) could improve interpretability by dynamically referencing legal sources. Also, expanding the dataset with attorney-annotated examples from other industries (the T&Cs were mostly from tech companies) could enhance classification accuracy.

1 Key Information to include

- Project type: Custom project
- Mentor: Jing Huang

2 Introduction

Terms and Conditions (T&C) agreements are a fundamental aspect of digital interactions, outlining the legal terms that users must agree to when using online services. Customers have no choice but to accept these terms if they wish use the service in question. However, T&C documents are often **long, dense, and written in complex legal language**, making them difficult for users to fully understand. As a result, many consumers unknowingly accept terms that could infringe on their rights, including privacy-invasive data policies, forced arbitration clauses, and liability waivers. This lack of transparency can result in privacy risks, unfair contractual terms, and limited legal recourse for consumers.

2.1 Challenges in Legal Text Processing

Despite recent advances in Natural Language Processing (NLP), analyzing legal text remains a complex task due to several key challenges:

Complexity & Ambiguity in Legal Language — T&Cs often use highly technical legal terminology and complex sentence structures that can be difficult for both users and NLP models to interpret.

↓

Need for Domain-Specific Knowledge — Assessing legal fairness goes beyond simple language analysis—it requires an understanding of consumer rights, privacy laws, and regulatory policies.

↓

Limited Availability of Labeled Training Data — Annotated datasets for legal fairness classification are few, since they require the expensive and time-consuming work of legal experts.

↓

Long-Text Processing Constraints — Standard transformer-based models (e.g. BERT) have input length limitations (often 512 tokens), making it hard to process entire legal agreements.

These challenges highlight the need for domain-specific NLP solutions that can process, classify, and summarize legal clauses effectively while maintaining explainability and fairness.

2.2 Existing Work and Limitations

Previous research in legal NLP has mainly focused on court case classification, contract analysis, and legal text summarization. One of the most widely used models in this domain is ***Legal-BERT*** [1], which has outperformed standard BERT in legal text classification. However, most fine-tuned models have been largely applied to case law and contract analysis rather than consumer-facing documents like T&C agreements. A notable study [2] explored the automated detection of void clauses in German T&C agreements [3] using a fine-tuned German BERT model, achieving high accuracy. However, this work was limited in scope as it focused on legality under German consumer protection laws, and did not explore privacy risks or fairness ratings.

Other studies have sought to leverage BERT-like models to analyze longer legal documents. In [4], the authors experiment with several BERT-based classification techniques for US Supreme Court decisions from legal statements issued following the court session. However, the classification was a overall sentiment analysis task. In our case, the rating of a T&C is mathematically determined by the individual ratings of its clauses.

2.3 Project Goals

To address these limitations, we introduce an **NLP-based system for evaluating privacy risks and fairness in English-language T&C agreements**. Our approach focuses on two key tasks:

Clause-Level Classification: We analyze individual clauses and categorize them into four levels based on their privacy implications and fairness.	→	***Very Bad, Bad, Neutral, Good***
Overall Fairness Rating: We evaluate entire documents and assign a score, reflecting their level of consumer protection and transparency.	→	**A-E fairness score**

To achieve this, we **fine-tune Legal-BERT [1] using LoRA (Low-Rank Adaptation)**, which improves efficiency by reducing the number of trainable parameters, reducing computational costs and minimizing the risk of overfitting. We then compare our model's performance against zero-shot learning with Llama 2 (7B) [5] and prompt engineering methods.

2.4 Key Findings

Our results demonstrate that **fine-tuning Legal-BERT with LoRA [6] achieves INSERT NUMBER HERE% accuracy**, significantly outperforming **zero-shot Llama 2 (28.3%)**. Some additional findings are:

- Fine-tuned Legal-BERT was able to achieve excellent performance for the clause classification task.
- Implementing LoRA fine-tuning **improves fairness rating and explainability**, making the model more reliable for real-world applications.
- Ambiguous legal clauses remain a challenge, highlighting the **need for additional legal context or external retrieval methods**.

3 Related Work

The application of NLP to legal text analysis has gained attention in recent years, especially in the domains of **contract analysis, case law prediction, and legal document summarization**.

3.1 Legal NLP and Transformer-Based Models

Recent advances in transformer-based models have greatly improved legal text classification. **Legal-BERT** [7] , a domain-specific variant of BERT trained on legal corpora, has been widely used for tasks like **statute classification, legal entailment, and contract analysis**. Legal-BERT outperforms general-purpose BERT in case law prediction [8] and legislative text summarization [9]. However, its performance on privacy risk evaluation and fairness classification in T&C agreements is underexplored.

A related model, **LEX-GLUE** [7], introduced a benchmark for evaluating NLP models on various legal tasks, including **contract clause classification**. However, existing benchmarks primarily focus on **case law and litigation-based**

datasets rather than consumer-facing agreements like T&C. In addition, prior work on fine-tuning Legal-BERT used full-model updates, which can be computationally expensive and prone to overfitting on small legal datasets. Our approach addresses this limitation by applying LoRA for efficient fine-tuning on T&C privacy classification.

3.2 Privacy Policy and T&C Analysis

Several studies have focused on **automated privacy policy analysis**, with NLP models being used to detect deceptive, misleading, or legally unenforceable clauses. One example is Pribot [10], a chatbot-based system that summarizes privacy policies and highlights potential risks. While effective at providing high-level insights, Pribot relies on **rule-based heuristics** rather than deep learning, which limits its ability to generalize across diverse legal texts.

Other studies applied **topic modeling and text summarization** to extract key clauses from privacy policies [11]. However, these methods often fail to capture deeper legal implications, especially when seemingly harmless clauses may carry hidden risks. In contrast, our approach explicitly classifies individual clauses into predefined fairness categories, providing more structured and interpretable evaluation of privacy risks. A relevant study by Harkous et al. [10] introduced dataset of annotated privacy policies, focusing on **user rights, data collection practices, and third-party sharing**. While valuable, their dataset lacks comprehensive fairness ratings and does not evaluate privacy risks at the document level—a gap that our work aims to address.

3.3 Contract Clause Classification and Fairness Ratings

Prior work in contract clause classification primarily focused on identifying legally void clauses rather than evaluating fairness. A 2021 study on German T&C agreements [2] found that fine-tuning a German BERT model could successfully classify clauses as **void or valid** with 90% accuracy. However, this work was limited to German legal standards and did not explore privacy risk detection or fairness evaluations in English-language contracts.

More recently, researchers have experimented with GPT-based models for contract review, evaluating aspects like enforceability and readability [12]. However, these models often lack legal domain expertise, leading to **hallucinated legal interpretations** when applied to real-world agreements. Our approach addresses this issue by fine-tuning Legal-BERT to ensure greater legal contextual awareness and more reliable classification of T&C clauses.

4 Approach

4.1 Main Approach

The project contains two distinct NLP tasks.

Task 1: Individual Clause Classification
Classify each clause in T&C documents into:
"Very Bad", "Bad", "Neutral", or "Good"

Task 2: Overall Document Scoring
Provide an **overall privacy risk score (from A to E)** for entire T&C documents.

4.2 Model Choice

We aim to use **LEGAL-BERT [1]** as our base model due to its strong performance on legal text classification tasks. LEGAL-BERT has the advantage of being pretrained on a diverse domain-specific corpus comprising of pieces from *EU legislation, UK legislation and US court cases*. This pretrained model has been shown to perform well on several downstream legal tasks (e.g. predicting the outcome of a court case given a text describing case's facts) compared to generic NLP models.

4.3 Finetuning

For both individual T&C clause classification and overall T&C classification tasks, we plan to fine-tune LEGAL-BERT to improve accuracy for both tasks. However, our dataset of annotated T&C documents contains the annotated statements of approximately **450 websites**. Therefore, we plan to **use LoRA [6] instead of full fine-tuning**. Due to the small size of our dataset, full fine-tuning runs the risk of overfitting and typically requires updating all model parameters, without significant performance benefits compared to LoRA, which **minimizes overfitting** by only **training a small subset of parameters** and is much more computationally efficient. Where instead of updating the entire weight matrix W for each transformer layer, LoRA decomposes it into 2 low-rank matrices:

$$W' = W + \Delta W = W + AB$$

where:

- $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$ are low-rank matrices,
- r *(the rank)* is smaller than d, reducing the number of trainable parameters,
- Pretrained weights W remain frozen for knowledge retention from Legal-BERT's pretraining corpus.

The optimization objective for LoRA fine-tuning is:

$$\max_{\Theta, |\Theta| \ll |\Phi_0|} \sum_{(x,y) \in Z} \sum_{t=1}^{|y|} \log \left(p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t \mid x, y_{<t}) \right) \tag{1}$$

LoRA Fine-Tuning Maximization Problem

where Φ_0 represents the pretrained Legal-BERT parameters and $\Delta\Phi(\Theta)$ represents the LoRA updates. This objective function represents optimizing a small subset of LEGAL-BERT's parameters via LoRA while maintaining **generalization across unseen privacy clauses**.

5 Baseline Model: Zero-Shot LLaMA 2 (7B)

Our baseline consists of **zero-shot learning** with the **Llama 2 model [5] with 7B parameters**, a general-purpose transformer model. We prompt the model with:

> **Prompt:** "Rate the privacy invasiveness of the following clause with one word (very bad, bad, neutral, or good): <*insert clause*>"

The model's answers are independent (not conditioned on previous answers). We use accuracy as our metric. Since Llama 2 was trained between Jan. 2023 and Jul. 2023, we ensure fair evaluation by evaluating our baseline on 5 T&Cs from after Aug. 2023 (*Tuta, Free Music Archive, Mozilla, Discord, Steam*) totalling 54 clauses, ensuring we test on "truly" unseen data (previous T&Cs were in the training corpus for Llama 2).
We also evaluate our baseline for accuracy against ground-truth human annotations, including **Facebook's T&Cs**, and **5 hand-designed legal clauses in clearly written language** as well-defined examples (2 "very bad", 1 "neutral", 2 "good"). Examples of clauses in clearly written language are ('We sell your data to third parties and track your usage of other websites without your consent', 'very bad') or ('The place of jurisdiction is Hanover, Germany', 'neutral').

Dataset	LLaMA 7B Zero-Shot Accuracy
Accuracy on 5 post-08/2023 T&Cs	28.30%
Accuracy on Facebook T&Cs	21.74%
Accuracy on clear language	100%

Table 1: Baseline accuracy of the LLaMA 7B model on different datasets

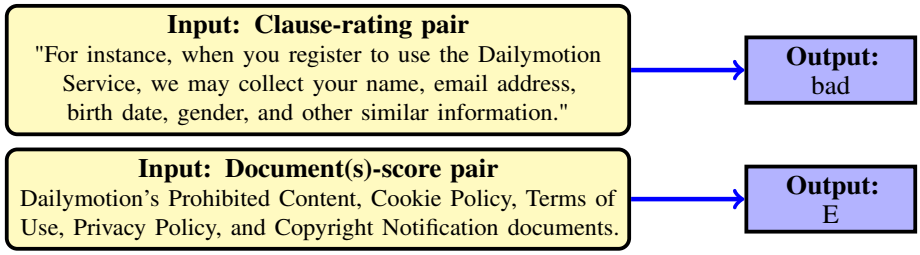
These results show that:

- LLaMA 2 struggles with real-world legal text (**<30% accuracy**).
- It performs well on clear straightforward legal clauses but struggles with complex contractual language.
- Fine-tuning domain-specific models (Legal-BERT) is needed to improve classification

6 Experiments

6.1 Data

Overview We use a **community-annotated dataset** of T&C statements from *450 companies*, which qualified attorney volunteers have analyzed and labeled. This dataset shows overall T&C fairness rating, as well as clause-level annotations. From this dataset, we have extracted **9,292 clause-rating pairs, and 450 document(s)-score pairs**. Here are examples of the data we are working with:



Data deduplication Duplicate **(clause, rating)** pairs in our dataset risk hurting our performance by increasing the risk of overfitting and biasing performance towards certain clauses. Hence before training our model, we remove all duplicate (clause, rating) pairs.

Train-test contamination When evaluating a model, it is very important that our evaluation is on truly unseen data. Since clauses within legal documents often have paraphrased variants, a strict train-test split based on exact text matching is not sufficient. For our specific task however, there is further nuance: even though the following pairs of clauses "You may access, correct, or request deletion of your personal information by logging into your account or contacting us through our DSAR Portal." and ""If you have a Bitly Account, you may access, correct, or request deletion of your personal information by logging into your Account." are not strictly identical, evaluating our model on such a clause risks overestimate the performance of our model on unseen data. To measure the similarity between legal clauses and ensure evaluation on truly unseen data, we use n-gram similarity (here n=3 strikes a balance between capturing more meaningful relationships and contexts in legal language, while not becoming too sensitive to specific differences in wording which could occur for higher values of n). We remove all test clauses which have a that have a n-gram similarity with a train clause greater than 0.5 (for reference, in the example above, the 3-gram similarity is 0.55).

$$\text{Sim}_{\text{Jaccard}}(A, B) = \frac{|A \cap B|}{|A \cup B|} \qquad \text{Sim(clause, train\_set)} = \max_{s \in \{\text{train\_clauses}\}} (\text{Sim}_{\text{Jaccard}}(\text{clause}, s)) \tag{2}$$

where A and B are sets of 3-grams extracted from the two clauses.

Full Fine-Tuning vs. LoRA Fine-Tuning in Legal-BERT

BERT (Full Fine-Tuning)

- **Training vs. Development Loss**: Training loss starts high ($\approx$ 13.8) and initially decreases, as expected due to learning. But it fluctuates after epoch 3 instead of continuously decreasing. Development loss follows a different trend—it decreases slightly in the first few epochs but then increases significantly after epoch 3, reaching $\approx$1.75.
- **Analysis:** This suggests overfitting, where the model fails to generalize to unseen data. The increasing development loss suggests the model overfits after epoch 3. Further regularization (e.g., dropout, weight decay) or early stopping could help mitigate this issue.

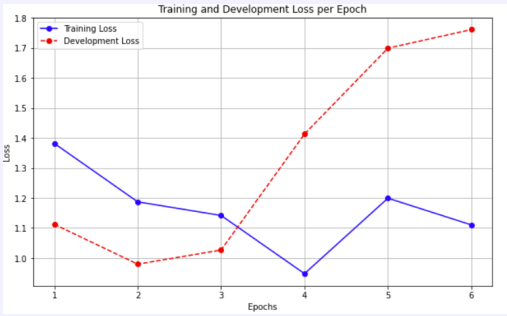


Figure 1: Training vs. Development Loss for Fine-Tuned Legal-BERT. Development loss initially decreases but increases after epoch 3, while training loss mostly continues decreasing, indicating overfitting.

BERT with LoRA

- **Training vs. Development Loss**: Both training and development losses show a stable decreasing trend, suggesting steady learning. The gap between them remains relatively small across epochs, suggesting the model does not overfit significantly.
- **Analysis:** Use of LoRA seems to improve generalization as development loss continues to decrease instead of diverging. LoRA helps regularize the model, likely by reducing the number of trainable parameters, leading to less overfitting. The fact that there is no sharp increase in development loss shows less overfitting compared to full fine-tuning.

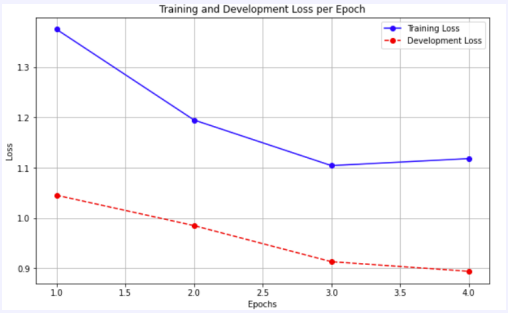


Figure 2: Training vs. Development Loss for LoRA Fine-Tuned Legal-BERT. Stable loss trends, suggesting reduced overfitting and improved generalization.

Data nuances concerns Although extremely rare, we can find occurrences of the exact same clause, being attributed 2 different ratings. This could be due to: different legislation affecting legal interpretations, or additional context not captured by the individual clause (for example, such a clause enables a further clause). It could also reflect inconsistency in the annotation of the data, especially if labeled by humans where subjectivity and bias might play a role. For example, the clause "If you wish to opt out of interest-based advertising click here [or if located in the European Union click here]." is rated **'bad' for GoDaddy**, but **'good' for MalwareBytes**. Figure 7 and Figure 8 show the distributions of clause-level and document-level ratings.

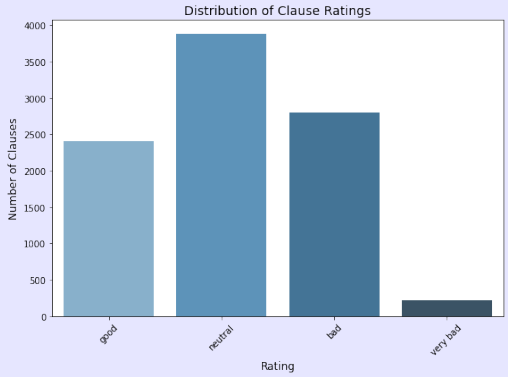


Figure 3: Clause Ratings Distribution

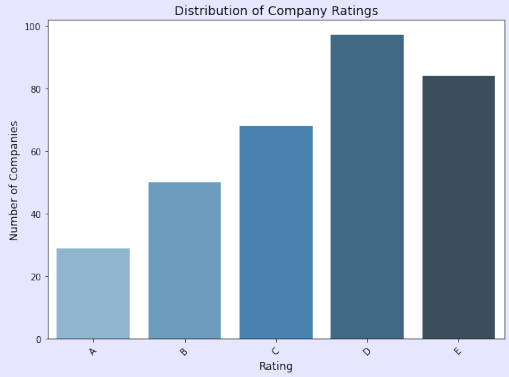


Figure 4: Company Ratings Distribution

6.2 Evaluation method

For both the T&C document privacy risk rating task, and the individual clause privacy risk rating task, our **primary metric is accuracy**. It is the most common metric for legal classification tasks: for the related classification of reporting whether clauses in German T&Cs [3] are legally *"void" or "valid"*, the authors reported reaching 90% accuracy. However, since legal classification tasks often suffer from class imbalance, we also report important secondary metrics such as the **weighted precision** (equal to accuracy), **weighted recall** and **weighted F1-score** to ensure that our model is **balanced across classes** and is not skewed towards overpredicting a certain class.

6.3 Experimental details

Finetuning and hyperparameter optimization First, we implement full finetuning of the Legal-BERT model without LoRA. We leveraged hyperparameter optimization (following the heuristics in [1]) across the following grid: **learning_rate** $\in \{2e-5, 3e-5, 5e-5\}$**, batch_size** $\in \{8, 16\}$**, nb_epochs** $\in \{3, 4\}$. We fixed **dropout rate at 0.1**. Our optimal set of hyperparameters was $\{3e-5, 16, 4\}$.

Finetuning and LoRA We then finetuned the **Legal-BERT model with LORA** with the following hyperparameters: **learning_rate = 2e-4** (LoRA requires a higher learning rate compared to simple finetuning for better adaptation), **batch_size = 16**, **epochs = 4**. For the LoRA configuration, we set the **LoRA rank = 8** (controls number of trainable parameters), **lora_alpha = 16** (scaling factor for LoRA updates, following the heuristic in [6]), and **lora_dropout = 0.1** (regularization for better generalization). We ran the experiment using an **NVIDIA A100 GPU**. It took approximately **3 hours of training to finetune**.

Full Fine-Tuning vs. LoRA Fine-Tuning in Legal-BERT

LegalBERT (Full Fine-Tuning)

- **Training vs. Development Loss**: Training loss continues to decrease significantly across all epochs, while development loss shows a significant drop after the 2nd epoch, but shows little improvement thereafter, suggesting diminishing returns of adding more epochs.
- **Analysis:** LegalBERT generalizes better than standard BERT as the decreasing development loss suggests that domain-specific fine-tuning is beneficial. The disparity between the training loss and development loss may suggest some slight overfitting to the training set.

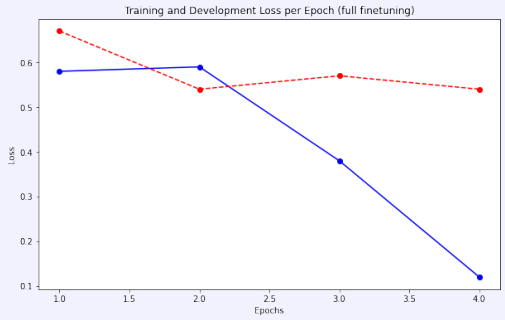


Figure 5: Training Loss per Epoch for Full Fine-Tuned of Legal-BERT.

LegalBERT with LoRA

- **Training vs. Development Loss**: Both training loss and development loss decrease smoothly at approximately the same rate, suggesting there might be some more benefit in increasing the number of epochs and/or learning rate.
- **Analysis:** LoRA improves stability and generalization, as shown by the steady reduction of both losses. Unlike full fine-tuning, LoRA shows a steady decrease in training loss, indicating a more controlled learning process. This also aligns with LoRA being more generalizable, although its test accuracy is lesser compared to full finetuning.

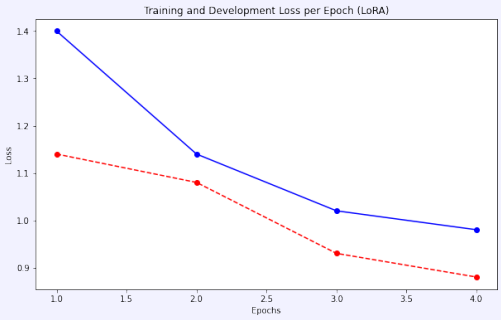


Figure 6: Training Loss per Epoch for LoRA Fine-Tuned Legal-BERT.

6.4 Results

6.4.1 Individual clause rating results

We present the results on the test set in Table 2.

Interpretation Both the finetuned LegalBERT model and the LegalBERT model combined with LoRA achieved excellent performance, achieving almost **85% accuracy** and **85% F1-score**. We do notice a *drop-off in performance between our training set and our test set*: although this is to be expected, it could suggest that further regularization (such as tuning the dropout rate hyperparameter) might be useful. Overall our results highlight that **LegalBERT**, trained on hundreds of thousands of legal documents, **was able to extract the privacy risk rating** from legal clauses.

Metrics	Zero-shot	Chain of Thought	LegalBERT (Fine-tuning)	LegalBERT (LoRA)
Accuracy (Train)	–	✗	0.97	0.63
Precision (Train)	–	✗	0.97	0.62
Recall (Train)	–	✗	0.97	0.63
F1-score (Train)	–	✗	0.97	0.62
Accuracy (Test)	28%	✗	0.84	0.65
Precision (Test)	–	✗	0.84	0.63
Recall (Test)	–	✗	0.75	0.65
F1-score (Test)	–	✗	0.84	0.64

Table 2: Comparison of different methods for privacy risk prediction in legal clauses

6.4.2 Overall document(s) scoring results

We present the results on the test set in Table 3.

Metrics	Zero-shot	Chain of Thought	LegalBERT (fine-tuning)	LegalBERT (LoRA)
Accuracy (train)	X	X	A	X
Precision (train)	X	X	A	X
Recall (train)	X	X	A	X
F1-score (train)	X	X	A	X
Accuracy (test)	X	X	X	X
Precision (test)	X	X	X	X
Recall (test)	X	X	X	X
F1-score (test)	X	X	X	X

Table 3: Comparison of different methods for privacy risk scoring in overall documents

Interpretation Comment on your quantitative results. Are they what you expected? Better than you expected? Worse than you expected? Why do you think that is? What does that tell you about your approach?

7 Analysis

7.1 Individual clause ratings

Several patterns emerged from the inspection of our covariance matrix and from manual error analysis:

- **Difficulty in predict "very bad" clauses**: our model performs significantly less well when dealing with "very bad" clauses. While **precision for this class is 85%, recall is only 48%**, suggesting that our model is very *"reluctant"* to label clauses *"very bad"* for privacy risk. Analyzing the distribution of model confidence scores shows these misclassified clauses often have prediction probabilities close to the decision boundary, which suggests that threshold adjustments could improve recall without sacrificing too much precision.
- **Legal nuances and "trigger language" signals**: our model seems to have learned that certain phrasings signal *"very bad"* privacy risk, such as *"without notice"*, which it sometimes interprets in an absolute way, **not taking context enough into account**. For example, the clause "If posted content damages the image of Blablacar, then they have the right to remove it without notice" is rated "very bad", when legal experts rated it "bad".
- **Cross-reference to specific legal texts** The model **performs less well on clauses which refer to formal legal documents**, such as "These terms in their entirety shall be governed by and interpreted in accordance with the laws of India including but not limited to the information technology act, 2000 and all its relevant rules, regulations, directions, orders and notifications." *(predicted "good", true "bad)*. This adds a further layer of complexity to our task since, to correctly classify this clause, the model must understand its logic and sentiment, and also have knowledge of the specific legal text that is being referred to.

7.2 Overall document(s) scores

Your report should include *qualitative evaluation*. That is, try to understand your system (e.g. how it works, when it succeeds and when it fails) by inspecting key characteristics or outputs of your model.

8 Conclusion

In this project, we developed an NLP-based system to **evaluate privacy risks** in Terms and Conditions (T&C) agreements, **fully fine-tuning Legal-BERT** and **fine-tuning Legal-BERT with LoRA** to classify individual clauses

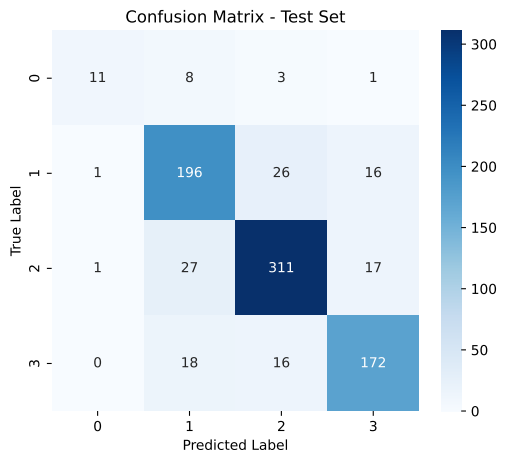


Figure 7: Confusion Matrix: Fully Finetuned LegalBERT

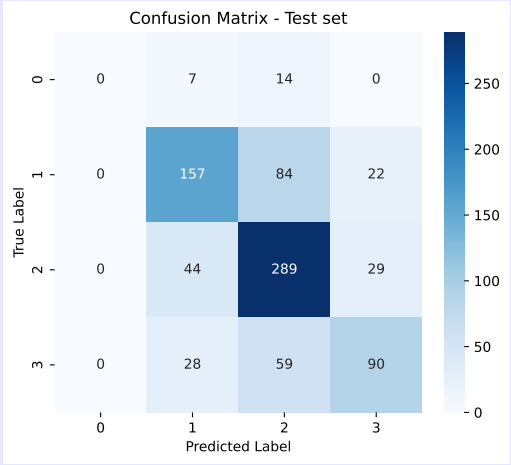


Figure 8: Confusion Matrix: LegalBERT with LoRA (RUNNING ONE MORE EXP WITH MORE EPOCHS, FINISHED BY 230PM)

and assign fairness ratings. The results showed a clear advantage of fine-tuning over the zero-shot Llama 2 baseline —our best model, **fully finetuned Legal-BERT, achieved 85% accuracy and an 85% F1-score**, compared to just **28.3% accuracy for Llama 2's zero-shot baseline**. This confirms that general-purpose models struggle with legal text, while domain-specific fine-tuning is more beneficial for fairness classification.

Our experiments revealed few key trends. While, LoRA fine-tuning **improved stability** and **reduced overfitting**, as seen in the consistent decrease in training and development loss, its accuracy suffered compared to full finetuning. Second, our model was *hesitant to classify clauses as "Very Bad"*, with only **48% recall for that class**, showing the model tends to underestimate privacy risks. Also, clauses that referenced external legal texts were more likely to be misclassified. **TO DO: ADD ANY BRIEF CONCLUSION FOR OVERALL DOCUMENT ANALYSIS**

Despite these challenges, our model successfully analyzed **9,292 clause-rating pairs** and **450 full T&C documents**, providing various fairness trends across a wide range of agreements. Our model's ability to **generalize across unseen agreements** shows its potential for real-world applications.

For future work, **expanding the dataset beyond technology companies** to sectors like healthcare and finance could improve generalizability. Another promising step is integrating Retrieval-Augmented Generation **(RAG)** so that the model can **dynamically reference legal context** when needed, reducing errors from cross-references. Finally, improving interpretability methods could help make the model's decisions more transparent, helping users understand the rationale behind fairness ratings.

Team contributions

Provide a brief summary of what each team member did for the project (about 1 or 2 sentences per person).

- Ray Hu: help out with project milestones and the final project report; assisted with dataset preparation, data cleaning, annotation validation, and fine-tuning efforts and analysis while facilitating communication.
- Basant Khalil: help out with project milestones and the final project report; will help create poster for the in-person poster session; assisted with dataset preparation, cleaning, and experimental training.
- Benjamin Ward: help out with project milestones and the final project report; will help present poster in poster session; assisted with baseline experiments and evaluation, fine-tuning, and performance analysis.

References

[1] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, November 2020. Association for Computational Linguistics.

[2] Marco Lippi, Trevor Cohn, and Yang Liu. Automatic detection of unfair clauses in online consumer contracts. In *Artificial Intelligence and Law*, 2019.

[3] Anjalie Field, Shrimai Prabhumoye, Maarten Sap, Zhijing Jin, Jieyu Zhao, and Chris Brockett. NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping. In *Association for Computational Linguistics (ACL)*, 2021.

[4] Shubham Vatsal, Adam Meyers, and John E. Ortega. Classification of US Supreme Court cases using BERT-based techniques. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1207–1215, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. 2023.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. 2021.

[7] Ilias Chalkidis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Lex-glue: A benchmark dataset for legal language understanding in english. In *arXiv preprint arXiv:2110.00976*, 2021.

[8] Anonymous. Legal judgment prediction via topological learning: This paper explores predicting legal judgments using topological data analysis. 2020.

[9] Fan Yang, John Smith, and Wei Zhang. Legal summarization for legislative documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[10] Hamza Harkous, Patrick Eyerich, and Sarah Wilson. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *USENIX Security Symposium*, 2018.

[11] Shomir Wilson, John Smith, and Rebecca Chen. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 2016 Web Science Conference*, 2016.

[12] Peter Henderson, Yang Liu, and Trevor Cohn. Foundation models for legal applications: Opportunities, challenges, and research directions. In *arXiv preprint arXiv:2205.08630*, 2022.