Can we teach the model twice?

Yifei Wang and Shaobo Yan

Abstract

Supervised fine-tuning has emerged as a common practice within the industrial domain. In this research project, our emphasis lies in the legal domain, where we delve into efficacious methods for fine-tuning large language models (LLMs) across diverse classification tasks, including but not limited to crime prediction and contract law classification. A comprehensive investigation was conducted across various parameter configurations and scenarios. The findings reveal that our model, denoted as the "lawyer" in this discourse, exhibits superior performance when subjected to fine-tuning on a comparable dataset. Such findings hold the potential to enhance the applicability of LLMs within the legal domain in the future. The codes are provided here.

1 Introduction

In contemporary industrial practices, supervised fine-tuning has emerged as a widely acknowledged and extensively employed methodology. To tailor language models to specific business requirements, developers often fine-tune pre-trained models, enabling them to be applicable to certain downstream tasks. When faced with new requirements, developers need to iterate on the model, leading to the following question: Should I use the previously fine-tuned model for old requirements, or start training from the original model?

In this project, our primary emphasis is on the legal domain, an active yet relatively unexplored area. Law interns and junior lawyers frequently encounter repetitive tasks that could be efficiently replaced by language models, adding significance to this research. Our focus lies in classification tasks involving full model fine-tuning, given that time efficiency is vital in the industry. Delving into various settings, we conducted comprehensive research to identify the most efficient approaches for accomplishing such tasks.

The report is structured as follows: Section 2 explains the datasets employed and the pre-trained Language Models (LLMs) we chose. Following this, Section 3 elaborates on the conducted experiments and visualizes the results through wandb. Subsequently, an in-depth exploration into the theoretical underpinnings of the observed curve is undertaken, accompanied by the presentation of conjectures. It is worth noting that further work is needed to figure out the math behind these phenomena.

2 Background

2.1 Models

- DeBERTa improves the BERT and RoBERTa models using disentangled attention and enhanced mask decoder. With those two improvements, DeBERTa out perform RoBERTa on a majority of NLU tasks with 80GB training data.(He et al., 2020)

- LEGAL-BERT is a family of BERT models for the legal domain, intended to assist legal NLP research. To pretrain the different variations of LEGAL-BERT, the authors collected 12 GB of diverse English legal text from several fields (e.g., legislation, court cases, contracts) scraped from publicly available resources. Sub-domain variants (CONTRACTS-, EURLEX-, ECHR-) and/or general LEGAL-BERT perform better than using BERT out of the box for domain-specific tasks.(Chalkidis et al., 2020)

2.2 Datasets

All of the datasets are taken from LexGLUE, which is a benchmark dataset to evaluate the performance of NLP methods in legal tasks. We select 3 subsets. (Chalkidis et al., 2021)

- Scotus: the US Supreme Court (SCOTUS) is the highest federal court in the United

States of America and generally hears only the most controversial or otherwise complex cases which have not been sufficiently well solved by lower courts. This is a single-label multi-class classification task, where given a document (court opinion), the task is to predict the relevant issue areas. The 14 issue areas cluster 278 issues whose focus is on the subject matter of the controversy (dispute).

- Ledgar: LEDGAR dataset aims contract provision (paragraph) classification. The contract provisions come from contracts obtained from the US Securities and Exchange Commission (SEC) filings, which are publicly available from EDGAR. Each label represents the single main topic (theme) of the corresponding contract provision. (Tuggener et al., 2020)

- Unfair-tos: the UNFAIR-ToS dataset contains 50 Terms of Service (ToS) from on-line platforms (e.g., YouTube, Ebay, Facebook, etc.). The dataset has been annotated on the sentence-level with 8 types of unfair contractual terms (sentences), meaning terms that potentially violate user rights according to the European consumer law.(Lippi et al., 2019)

3 Experiments

In this section, we will provide a concise overview of the conducted experiments.

3.1 Settings

The foundation of our experiments is grounded in five distinct research questions. The first question simply considers the most general setting, exploring the efficiency comparison between two approaches. Furthermore, we delve into the impact of substantial variations in the sizes of two datasets. The role of model type in influencing outcomes is also investigated. Additionally, we examine the significance of labels and explore whether fine-tuning on an unrelated dataset yields consistent results. Bearing these considerations in mind, we designed five experiments leveraging the aforementioned models and datasets. To minimized the impact of randomness, each experiment was ran three times with fixed seeds, and the result curve is shown through taking the mean values of the results. The hyperparameters used in these experiments are listed in Table 1.

Parameter	Value
`epoch`	5
`learning rate`	1e-5
`max input length`	512
`batch size`	32
`seed`	111 222 333

Table 1: Hyperparameters

3.2 Result

Given the utilization of datasets A and B in the experiment, for the sake of simplicity, we exclusively present the results evaluated on dataset B, thereby simplify the report. Figures 1-5 visualize the outcomes, and a comprehensive result will be conducted in the subsequent section. Our evaluation metrics includes accuracy, micro f1 and macro f1. Loss curve is also taken into consideration.

4 Conclusion

4.1 Discussion

The analysis above provides insights into how to approach such tasks in real-life scenarios.If you find yourself, as a law intern, tasked with such an assignment by your mentor, the initial step is to figure out whether there already exists a fine-tuned model and the dataset on which it has been trained. Utilizing a fine-tuned model trained on a dataset with similar content and size undoubtedly enhances our operational efficiency. It is advisable to start from scratch if the two datasets are unrelated or if there is a significant difference in quantity. Additionally, the selection of the appropriate pretrained model also plays a crucial role.

4.2 Theoretical explanation

A fine-tuned model exhibits a smoother distribution of the data labels in contrast to the original Language Model (LLM). This distinction arises from the fact that the latter one utilizes a randomly initialized header to do predictions in the beginning. Consequently, when both models are trained for an equivalent duration, the fine-tuned model initially outperforms and tends to sustain this advantage. However, when the model has been tuned on a rather small dataset, the distribution is peak when transferring to do predictions on a large dataset, which can be interpreted as a bad initialization, potentially leading to sub-optimal performance, as illustrated in the figure 6.

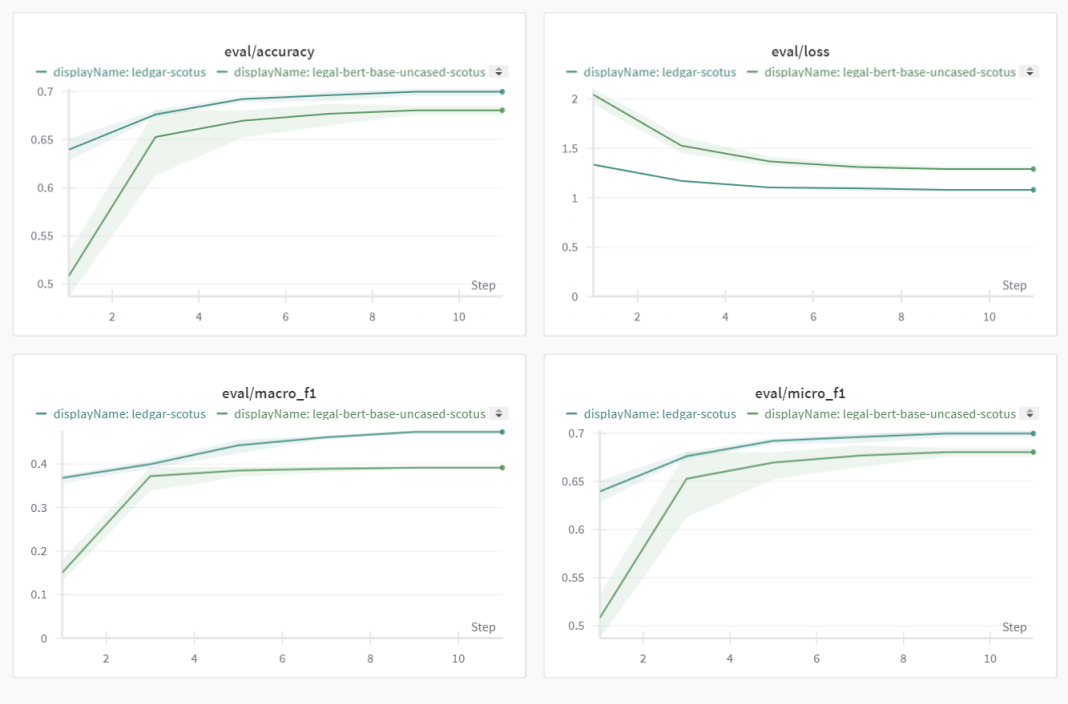

Figure 1: Predictions on ledgar dataset. This is the original problem setting. The upper line is the curve of model fine-tuned on scotus, while the lower line show the result of the original LEGAL-BERT. The formal method achieves better accuracy at the beginning and performs better in each epoch's evaluation.

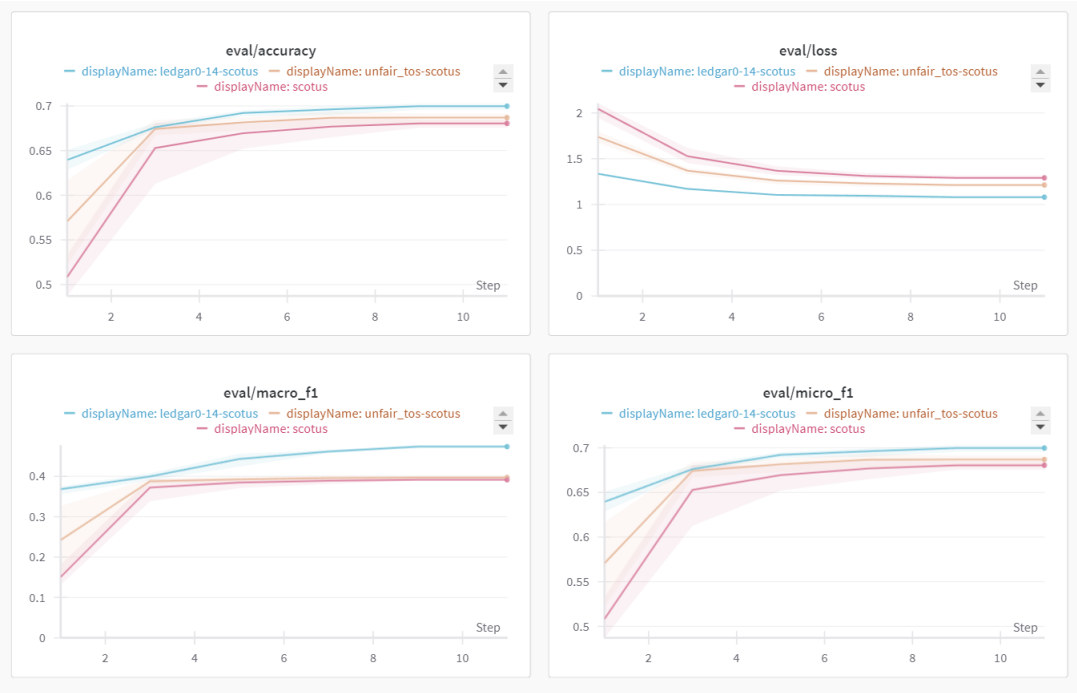

Figure 2: Predictions on scotus dataset. In this experiment, we introduced a third dataset, namely unfair-tos, anticipated to exhibit lesser similarity to scotus in comparison to ledgar. The upper curve represents the performance of the model fine-tuned on ledgar, the middle curve illustrates the outcomes of the model tuned on unfair-tos, and the lower curve depicts the results of the original LEGAL-BERT. The fine-tuned method consistently demonstrates superior performance compared to the original model, irrespective of the dataset utilized. Nevertheless, it is noteworthy that the similarity between datasets does exert an impact on performance, with closer similarities yielding better results.

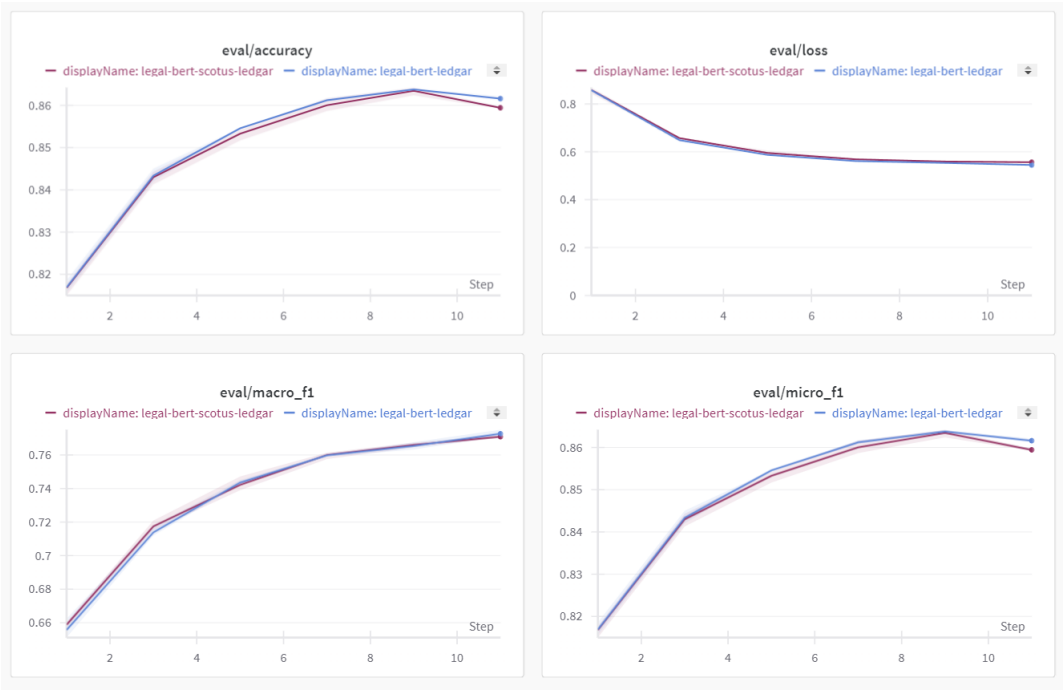

Figure 3: Predictions on ledgar dataset. Ledgar is 100 times larger than scotus. The lower line is the curve of model fine-tuned on scotus, while the upper line show the result of the original LEGAL-BERT. It is evident from the results that fine-tuning on a small dataset does not lead to performance improvement.

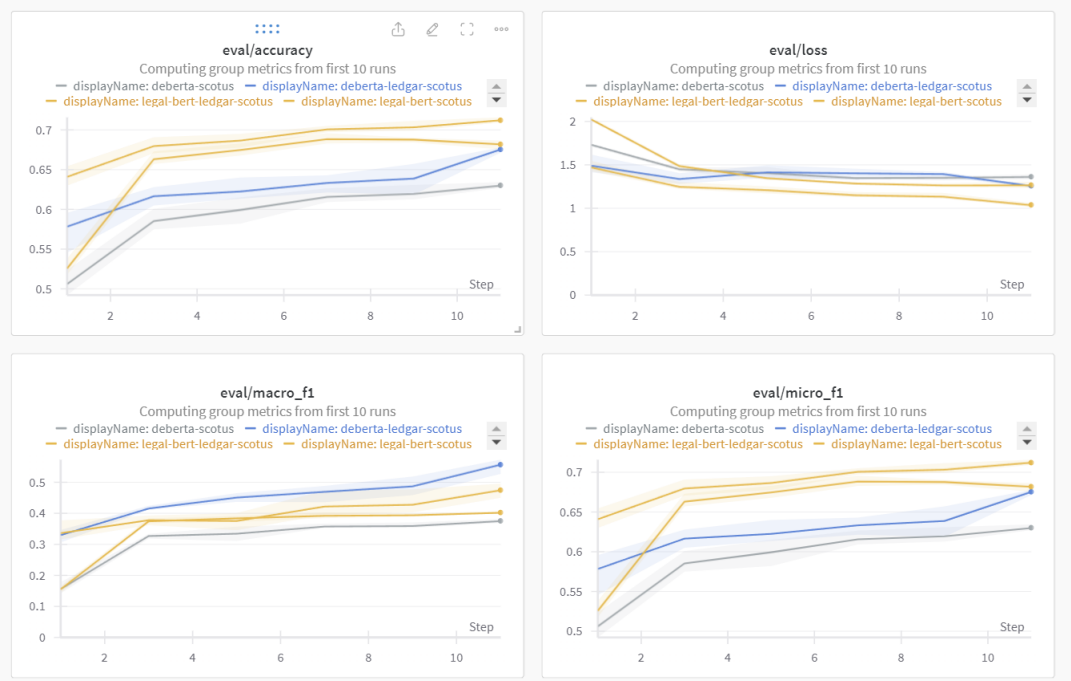

Figure 4: Similar to the previous experiment, changing the model type while keeping other factors constant results in analogous behaviors.

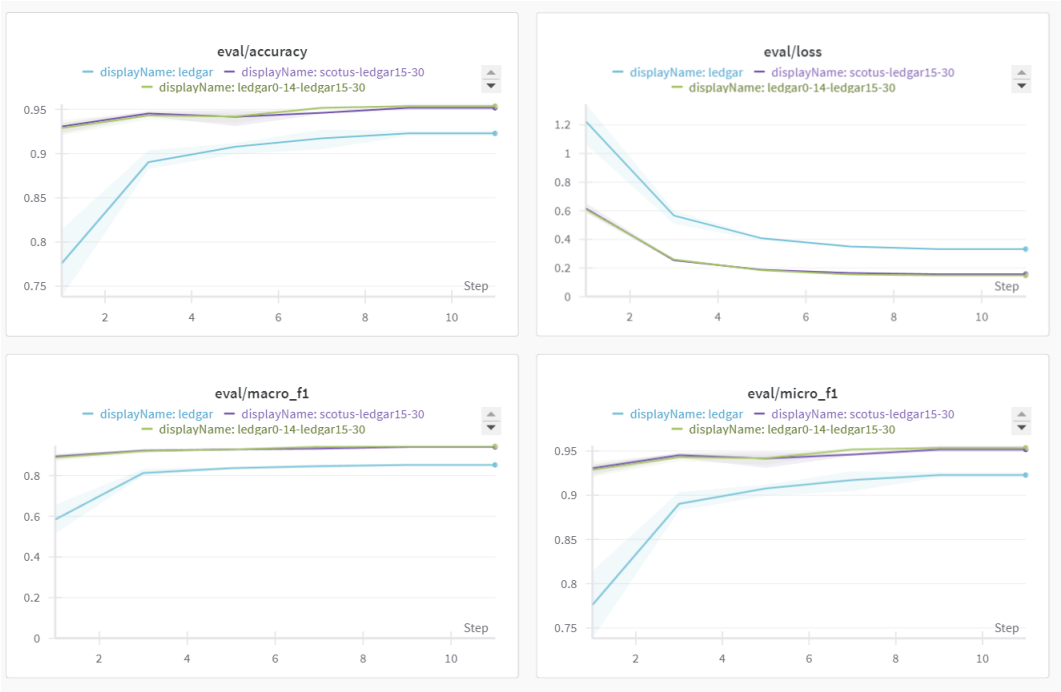

Figure 5: Predictions on ledgar dataset. In this experiment, we selected half of the label types from the ledgar dataset and utilized the remaining label types for predictions. Remarkably, the models fine-tuned on scotus and ledgar demonstrated similar results.

Nevertheless, the aforementioned analysis remains speculative, and the establishment of a solid mathematical formulation is left for future research.

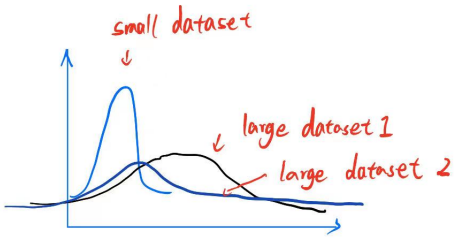

Figure 6: Distributions given by the model

5 Ackonwledgement

The authors express gratitude to the Teaching Assistants and the instructor for their constructive feedback. Additionally, the author acknowledges the valuable discussions with friends majoring in law, which proved to be instrumental and highly appreciated in the development of this project.

6 Contributions

The authors contributed equally.

Yifei Wang wrote the main training code and the report, design the shell script for running the experiments and run part of the experiments.

Shaobo Yan wrote the helper function file, implemented our own version of datasets, prepared the PPT for presentation, and run part of the experiments.

References

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, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139.

Don Tuggener, Pius Von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. Ledgar: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241.