Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review

Fusheng Wei
Data & Technology
Ankura Consulting Group, LLC
Washington, D.C. USA
fusheng.wei@ankura.com

Robert Keeling, Esq.
Complex Commercial Litigation
Sidley Austin LLP
Washington DC, USA
rkeeling@sidley.com

Nathaniel Huber-Fliflet
Data & Technology
Ankura Consulting Group, LLC
Washington DC, USA
nathaniel.huber-flilet@ankura.com

Jianping Zhang
Data & Technology
Ankura Consulting Group, LLC
Washington DC, USA
jianping.zhang@ankura.com

Adam Dabrowski
Data & Technology
Ankura Consulting Group LLC
Washington, D.C. USA
adam.dabrowski@ankura.com

Jingchao Yang
Data & Technology
Ankura Consulting Group, LLC
Washington, D.C. USA
jingchao.yang@ankura.com

Qiang Mao
Data & Technology
Ankura Consulting Group, LLC
Washington, D.C. USA
qiang.mao@ankura.com

Han Qin
Data & Technology
Ankura Consulting Group LLC
Washington, D.C. USA
han.qin@ankura.com

Abstract—**The increased integration of Large Language Models (LLMs) across industry sectors is enabling domain experts with new text classification optimization methods. These LLMs are pretrained on exceedingly large amounts of data; however, practitioners can perform additional training, or "fine-tuning," to improve their text classifier's results for their own use cases. This paper presents a series of experiments comparing a standard, pretrained DistilBERT model and a fine-tuned DistilBERT model, both leveraged for the downstream NLP task of text classification. Tuning the model using domain-specific data from real-world legal matters suggests fine-tuning improves the performance of LLM text classifiers.**

To evaluate the performance of text classification models, using these two Large Language Models, we employed two distinct approaches that 1) score a whole document's text for prediction and 2) score snippets (sentence-level components of a document) of text for prediction. When comparing the two approaches we found that one prediction method outperforms the other, depending on the use case.

Keywords—LLM, MLM, fine-tuning, text classification, large language model, predictive modeling, TAR, predictive coding

I. INTRODUCTION

With recent advancements in Large Language Models (LLMs), it has become imperative for downstream industries to identify applications of LLMs within each business domain. Among innovative industries, the legal industry is one at the forefront of this pursuit, given its common practice of applying predictive modeling – known in the legal industry as 'predictive coding' or 'Technology Assisted Review (TAR)' – which is a popular tool used to augment a manual document review and the text classification process.

Initially, the integration of machine learning in legal disputes has involved traditional methods like Logistic Regression (LR) and Support Vector Machine (SVM). Recent developments in machine learning and artificial intelligence have expedited the need to incorporate deep learning into the TAR toolkit. As LLMs continue evolving into state-of-the-art deep learning methods, it naturally becomes viable and imperative to explore their applications in legal disputes.

Currently, there are two prominent architectures of LLMs: Masked Language Models (MLM) and Causal Language Models (CLM). BERT, and permutations of this model, represent the MLMs, while GPT and other generative models represent the latter. MLMs and CLMs are both built upon transformer architecture, which represents the foundation of Large Language Models. When using MLMs, the model is trained to predict masked tokens within the input sequence (e.g., a sentence). Whereas in CLMs, the model is trained to predict the next token in the input sequence.

While both types of models handle Natural Language Processing (NLP) tasks, their functions and use cases can vary widely. MLMs are frequently utilized for tasks such as text classification, sentiment analysis, and named entity recognition. Alternatively, CLMs specialize in tasks like text generation and summarization.

LLMs are initially pre-trained on extensive generic data, like BERT being pre-trained on Wikipedia and Google's BooksCorpus data. Training an LLM with such copious data makes the model extremely robust, but simultaneously makes the model universal and not attuned to any specific domain. Additional training of an LLM, or "fine-tuning", is critical to align the model with a specific use case to improve results.

979-8-3503-2445-7/23/$31.00 ©2023 IEEE

Prominent pretrained LLMs can be fine-tuned and applied to specific tasks, such as text classification.

Fine-tuning an LLM leverages a set of domain-specific text exemplars as additional training data for the existing, pretrained model. Tuning is considered self-supervised learning, where words are masked randomly and used as labels to retrain a small set of parameters within the original model. This tuning method normally does not require human-labeled training data. Fine-tuned LLMs, using human-labeled training sets and applied in a text classification scenario, are becoming more popular in the legal domain. Generally, it is believed that this approach develops effective models, but concurrently alludes to the potential benefits of fine-tuning the LLM before implementing it for text classification .

There is the potential for performance improvements in NLP tasks when a fine-tuned LLM is used for text classification. The fine-tuning process acclimates the underlying LLM to the unique characteristics and nuances of the textual data within the domain-specific data before it is used for classification. While applying LLMs that have not been fine-tuned to text classification tasks often yields acceptable performance, experimenting with fine-tuned LLMs provides a compelling opportunity to further improve performance in certain NLP applications.

In this paper, we conducted a series of experiments that o examined the performance impact of fine-tuning an LLM (DistilBERT) in a text classification scenario. The experiments were conducted using three data sets from confidential, non-public, real-world legal matters across various industries. These data sets were comprised of unstructured data, including emails and other electronic document types such as Microsoft Office, PDFs, and text files. A subset of this data was used to fine-tune a pretrained DistilBERT model – the model was then applied to a text classification task for each matter's data set. In our experiments, we compared a fine-tuned DistilBERT model to an "out of the box" pretrained DistilBERT model.

Our experiments demonstrate that the fine-tuned DistilBERT model consistently outperforms the "out of the box" pretrained DistilBERT model when applied to text classification. This observation underscores the importance of incorporating domain-specific data into an LLM's fine-tuning for its subsequent deployment on a text classification task.

To assess the performance of fine-tuning, we used two distinct approaches. First, we applied each model to classify a document's entire text and second, we applied each model to classify only snippets of text from the same document set. A snippet, in our experiments, is a component part of a document's text, typically two or three sentences. We found that fine-tuning DistilBERT performed demonstrably better than the "out of the box" pretrained DistilBERT model. Additionally, when comparing the document-level and snippet-level results of the fine-tuned model, snippet classification outperformed document classification when applied to one data set, while document classification yielded superior results when applied to the other two data sets.

Finally, we compared the performance of the fine-tuned DistilBERT model with a traditional Logistic Regression model. Our findings indicate that both approaches – fine-tuned LLMs and traditional Logistic Regression models – provide effective solutions for text classification in legal matters. The versatility of these distinct approaches suggests potentially applying new modeling strategies to improve the performance of text classification tasks in the legal domain.

Prior publications of text classification research in the legal domain are discussed in Section II. Section III details the experiment methodology and construction. The experimental results are presented in Section IV, and our findings and conclusions are summarized in Section V.

II. Related Work

Machine learning techniques, such as text classification are well established in the legal domain with Logistic Regression and Support Vector Machine being two popular machine learning algorithms for this task [1]. These algorithms learn from features generated by tokenization from bag of words. Before applying transformer models to text classification, prior studies applied deep learning methods, such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LTSM), for text classification in legal document review [2, 3, 4]. CNN demonstrated good performance, yet across various data sets no single algorithm consistently outperformed the others.

In recent years, LLMs have surpassed deep learning models as the state-of-the-art architecture in every NLP aspect. Created by Google in 2018 [5], BERT is an extremely prevalent LLM that allows for transfer learning in NLP tasks by fine-tuning "out of the box" pretrained models on domain-specific data for downstream applications. Zhao, Ye and Yang (2021) [6] studied the effectiveness of transfer learning using BERT in privileged document review when compared to Logistic Regression for the same text classification task. The experiments yielded mixed performance improvement results.

In "An Empirical Comparison of DistilBERT, Longformer and Logistic Regression for Predictive Coding" [7], the authors tested the application of DistilBERT and Longformer in text classification. The results demonstrated that Longformer performs better or similar to DistilBERT and Logistic Regression because Longformer can handle more tokens as input compared to the other algorithms. However, due to Longformer's training and compute time, it is not practical to use with real-world document review projects. This study also briefly tested fine-tuning the DistilBERT model with domain-specific data. In Wei *et al.* (2022) [7], the LLM was fine-tuned with publicly available legal domain data and used to measure its performance in an active learning text classification task.

Using text classification to classify snippets of text from documents is gaining popularity, especially in the legal domain [8, 9]. In this approach, documents are broken into snippets, a small passage of words usually ranging from 50 to 200 words, and the model is applied to all snippets from the document. The highest scoring snippet then represents the score for the whole document. Snippet classification augments Explainable AI and simplifies the explanation of why the model made its classification decision, further minimizing the black box nature of text classification.

III. Experiments

A. Data Sets

Our experiments were conducted on three data sets from confidential, non-public, real-world legal matters across various industries. These data sets were comprised of unstructured data, including emails and other electronic document types, such as Microsoft Office, PDFs, and text files. We reduced the data size for fine-tuning to improve the speed of the process and to avoid overfitting. The data for fine-tuning was limited by removing file types with large quantities of text and files that may contain unhelpful textual structure or patterns, such as Microsoft Excel files. The filtered fine-tuning data was further cleansed by removing email headers, email footers, URLs, and duplicative text. Table I provides the breakdown of fine-tuning data for each of the three data sets.

TABLE I. Data sets for MLM Fine-Tuning

Matter	**Total Number of Documents**	**Number of Documents Used for Fine-Tuning**
Project A	4,000,000	400,000
Project B	1,000,000	300,000
Project C	800,000	250,000

Table II provides the breakdown of the labels for the three data sets for the text classification experiments.

TABLE II. Data sets for Text Classification

Project	**Document Coding Type**	**Number of Training Documents**	**Percentage of Positive Training Documents**	**Number of Testing Documents**	**Percentage of Positive Testing Documents**
Project A	Responsive	10,437	17.34%	17,284	1.36%
Project B	Privilege	18,372	20%	4,593	20%
Project C	Privilege	18,789	20%	5,000	12%

B. Methodology

These experiments used the DistilBERT LLM [10] downloaded from Hugging Face and included checkpoint 'distilbert-base-uncased' for the fine-tuning process. This model has 66 million parameters making it a distilled version of the BERT LLM. Hugging Face's API was used for both the language model and text classification fine-tuning.

Language model fine-tuning begins with tokenizing training text documents and then concatenating and breaking the tokenized vectors into 512 size chunks. These chunks are 15% masked for self-supervised tuning. The training output consists of adjusted parameters, a tuned tokenizer, and a vocabulary word list.

Once the LLM fine-tuning completes, then the text classification models are created. Two text classification models were created for each of the three data sets: a model using the fine-tuned DistilBERT LLM and one using the standard, pretrained DistilBERT LLM.

The standard, pretrained DistilBERT LLM and the fine-tuned DistilBERT LLM are then used to score the testing data sets of each project. The first text classification task applies the models to whole documents by ingesting the first 512 tokens of text to make the prediction. This value is the default token limit for DistilBERT. The second text classification task breaks the documents into snippets of text and applies the models to these snippets. The highest scoring snippet from each document then represents the whole document's score.

Lastly, a Logistic Regression model is created for each matter and tested using whole documents and snippets allowing for comparison to the DistilBERT LLMs. In summation, each of the data sets were assessed by three models and had three sets of scores.

The computation was executed on an Azure cloud server equipped with four A100 GPU cards.

C. Evaluation Metrics

Our key performance metrics measured the overall precision, recall, and F1 score of each of the models against each test population at a set recall of 75%. This recall level was chosen as a midpoint between the generally accepted range of 70% - 80% recall in most legal domain text classification projects. Precision-recall curves were also used to visualize and measure each model's performance.

IV. Results and Discussion

A. A Peek Into the Fine-Tuned LLM

Before analyzing the results of the text classification experiments, we examined the effect fine-tuning had on the DistilBERT LLM. As mentioned previously, LLM fine-tuning is a type of domain adaptation to help the model better understand domain-specific language. The following results (*Example 1 and Example 2*) from Project C's fine-tuned model show that the model adapts well when fine-tuned using legal domain data. The "Original Text" in the examples are from real-world data with their confidential information removed.

In the examples, a word in the "Original Text" is masked from the standard pretrained DistilBERT model and also from the fine-tuned DistilBERT model. In this case, the models' job is to predict the masked word. The masked words in these examples are "*legal*" and "*truth*," respectively. Interestingly, the fine-tuned DistilBERT model was able to predict both masked words while the standard pretrained model couldn't predict either masked word.

Example 1:

Original text: *Will defer to legal on approvals for the next step.*
Masked text: *Will defer to [**MASK**] on approvals for the next step.*

The five highest scoring predictions from each model are listed below:

Pretrained DistilBERT Model:

>>> Will defer to vote on approvals for the next step.
>>> Will defer to comment on approvals for the next step.
>>> Will defer to rely on approvals for the next step.
>>> Will defer to agree on approvals for the next step.
>>> Will defer to depend on approvals for the next step.

Fine-Tuned DistilBERT model
*>>> Will defer to **legal** on approvals for the next step.*
>>> Will defer to Dave on approvals for the next step.
>>> Will defer to them on approvals for the next step.
>>> Will defer to leadership on approvals for the next step.
>>> Will defer to you on approvals for the next step.

Example 2:

Original text: *We used the ground truth of sales figures from last year as a basis for forecasting this year's results.*
Masked text: *We used the ground [**MASK**] of sales figures from last year as a basis for forecasting this year's results.*

The five highest scoring predictions from each model are listed below:

Pretrained DistilBERT Model:

>>> We used the ground ##work of sales figures from last year as a basi s for forecasting this year's results.
>>> We used the ground breaking of sales figures from last year as a ba sis for forecasting this year's results.
>>> We used the ground ##ings of sales figures from last year as a basis for forecasting this year's results.'
>>> We used the ground ##ing of sales figures from last year as a basis for forecasting this year's results.'
>>> We used the ground zero of sales figures from last year as a basis fo r forecasting this year's results.'

Fine-Tuned DistilBERT Model
*>>> We used the ground **truth** of sales figures from last year as a basis f or forecasting this year's results.*
>>> We used the ground data of sales figures from last year as a basis f or forecasting this year's results.
>>> We used the ground breaking of sales figures from last year as a ba sis for forecasting this year's results.
>>> We used the ground share of sales figures from last year as a basis f or forecasting this year's results.
>>> We used the ground up of sales figures from last year as a basis for forecasting this year's results.

B. *Empirically Comparing the Pretrained and Fine-Tuned Models*

Charting the precision-recall curves for each of the model outputs across the three projects, we can visually compare and assess model performance.

Figures 1-3 show the comparison of the pretrained DistilBERT model to the fine-tuned DistilBERT model at the document level. The figures illustrate that the fine-tuned DistilBERT model yields better results than the pretrained DistilBERT model. Notably, at higher recall rates, both models perform similarly for Project B and Project C.

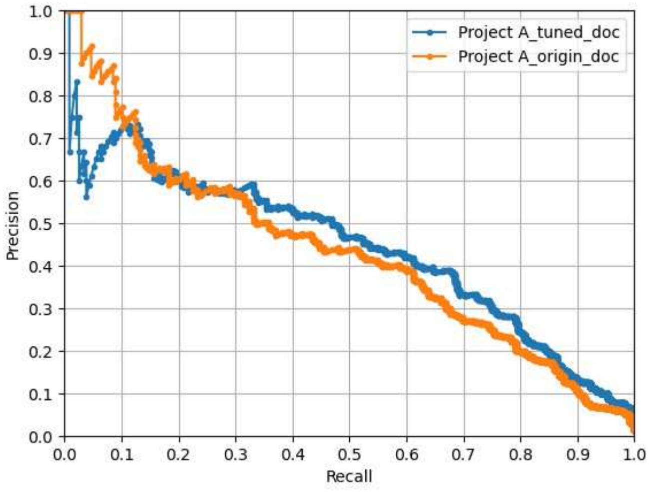

Fig. 1. Project A – Precision-recall curves for pretrained DistilBERT model and fine-tuned DistilBERT model on whole documents.

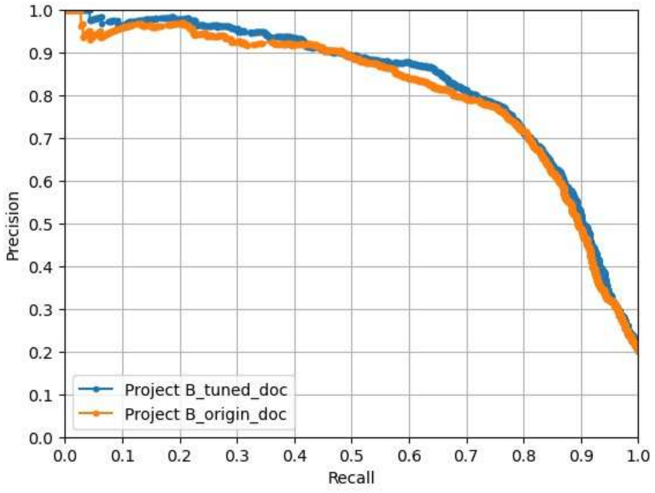

Fig. 2. Project B – Precision-recall curves for pretrained DistilBERT model and fine-tuned DistilBERT model on whole documents.

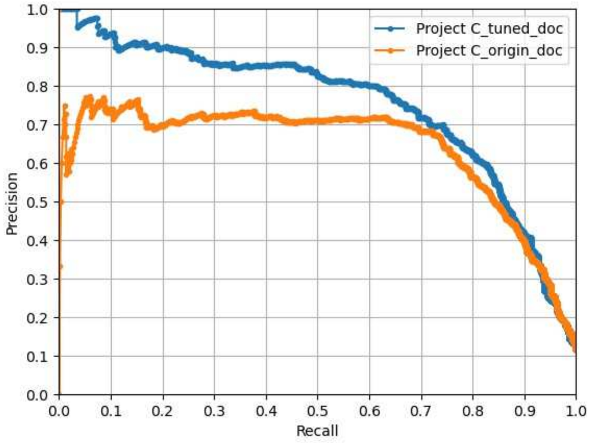

Fig. 3. Project C – Precision-recall curves for pretrained DistilBERT model and fine-tuned DistilBERT model on whole documents.

Figures 4-6 show the comparison of the pretrained DistilBERT model to the fine-tuned DistilBERT model at the snippet level. To serve as an additional benchmark for comparison, the performance of the fine-tuned DistilBERT model at the document level is plotted on the same graph. While the fine-tuned DistilBERT document model still performed the best for Project C, the fine-tuned DistilBERT snippet model performed the best for Project A. While Project B presents mixed results because the fine-tuned DistilBERT document model is generally better than the fine-tuned DistilBERT snippet model, however, the fine-tuned DistilBERT snippet model has higher precision at very high recall rates.

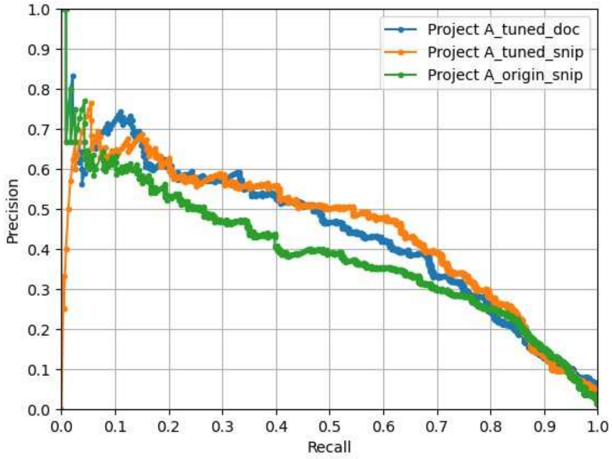

Fig. 4. Project A – Precision-recall curves for pretrained DistilBERT model and fine-tuned DistilBERT model on snippets of text and fine-tuned DistilBERT model on whole documents.

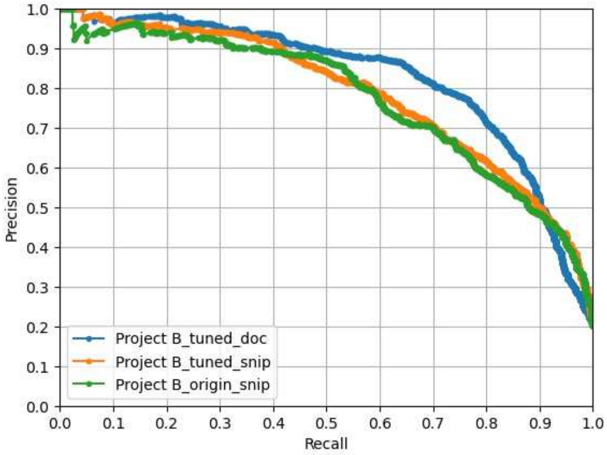

Fig. 5. Project B – Precision-recall curves for pretrained DistilBERT model and fine-tuned DistilBERT model on snippets of text and fine-tuned DistilBERT model on whole documents.

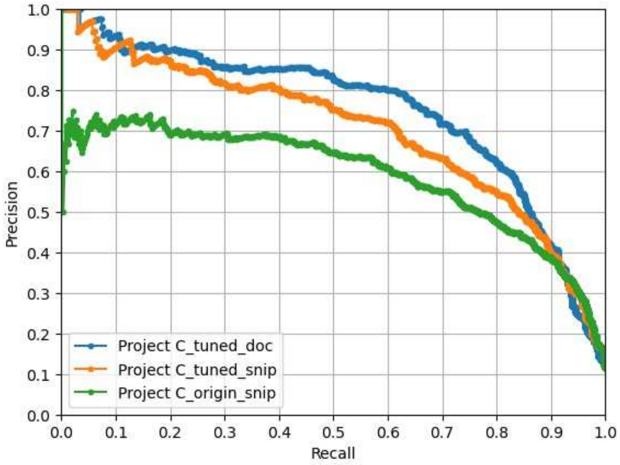

Fig. 6. Project C – Precision-recall curves for pretrained DistilBERT model and fine-tuned DistilBERT model on snippets of text and fine-tuned DistilBERT model on whole documents.

Figures 7-9 compare the best performing DistilBERT LLM for each project from the prior comparisons with the Logistic Regression model. Logistic Regression models are well used in legal document classification and provide another helpful assessment. The Logistic Regression model was assessed at a document and snippet level. Project A compares the fine-tuned DistilBERT snippet model with the Logistic Regression model, while Project B and Project C compare the fine-tuned DistilBERT document model with the Logistic Regression model. The figures illustrate that the fine-tuned DistilBERT model tends to perform better than or similarly to the Logistic Regression model for all projects at 75% recall.

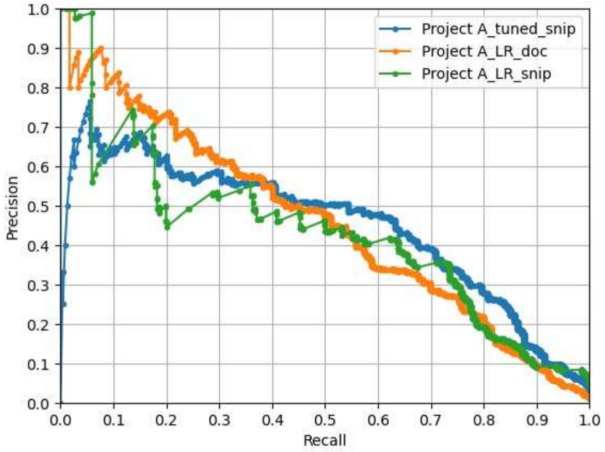

Fig. 7. Project A – Precision-recall curves for fine-tuned DistilBERT model on snippets of text and logistic regression models on whole documents and snippets of text.

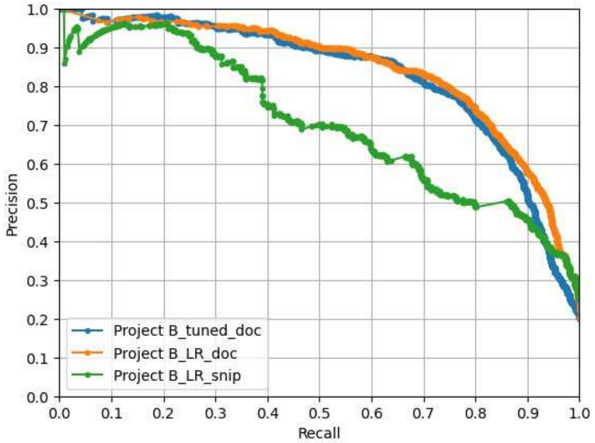

Fig. 8. Project B – Precision-recall curves for fine-tuned DistilBERT model on whole documents and logistic regression models on whole documents and snippets of text.

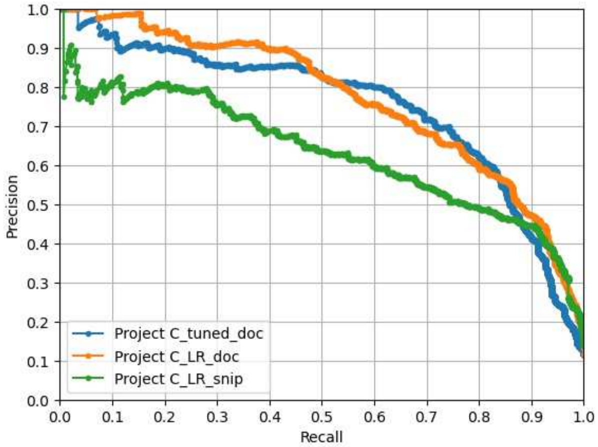

Fig. 9. Project C – Precision-recall curves for fine-tuned DistilBERT model on whole documents and logistic regression models on whole documents and snippets of text.

Table III provides the precision and F1 values for each of the model outputs across the three projects. Examining both precision and F1 scores at 75% recall for each project, performance shows that the fine-tuned DistilBERT models, document and snippet, perform better than the other models.

TABLE III. PRECISION AT 75% RECALL

Matter	Model	Precison	F1
Project A	Project A_tuned_doc	29.65%	42.55%
	Project A tuned_snip	**33.78%**	46.62%
	Project A_origin_doc	24.89%	37.42%
	Project A_origin_snip	28.43%	41.27%
	Project A_LR_doc	26.74%	39.46%
	Project A_LR_snip	29.96%	42.85%
Project B	Project B_tuned_doc	**77.99%**	76.49%
	Project B tuned_snip	65.66%	70.05%
	Project B_origin_doc	77.38%	76.20%
	Project B_origin_snip	64.51%	69.38%
	Project B_LR_doc	79.59%	77.26%
	Project B_LR_snip	51.97%	61.66%
Project C	Project C_tuned_doc	**67.50%**	71.09%
	Project C tuned_snip	58.73%	65.91%
	Project C_origin_doc	63.82%	69.00%
	Project C_origin_snip	51.06%	60.78%
	Project C_LR_doc	65.36%	69.89%
	Project C_LR_snip	50.70%	60.55%

V. CONCLUSIONS

LMM research is currently very popular in the legal domain. Small improvements in the precision of a text classification model can have significant reductions in the cost and risk of a legal document review. The goal of this research was to measure the performance of fine-tuned LLMs in legal document review. Our work provides an empirical assessment of fine-tuned LLMs that legal practitioners can use to influence their design of text classification modeling strategies.

The experiments show that fine-tuning the LLM can improve performance of subsequent text classification. The results also show that, depending on the project, text classification with a fine-tuned LLM applied at a snippet-level can perform better than document-level classification. Regardless of the choice of document segmentation, the fine-tuned LLM always performs better than the standard pretrained version. Lastly, Logistic Regression models perform well at a variety of recall rates when compare to the fine-tuned LLM suggesting there is still a prominent place Logistic Regression should have in text classification for legal document review.

It is important to note that LLM fine-tuning requires investment in GPU infrastructure and takes considerable time for training.

In our experiments we used Hugging Face's default tuning parameters and in future work we plan to explore customized settings, especially controlling the layers of the tuning parameters.

The results of the fine-tuned DistilBERT models applied to snippet text continues to drive research into the effectiveness this scoring strategy in legal document classification. In prior work, the authors observed that text classification of snippets can sometimes yield higher precision, which saves cost and reduces risk in legal document review. We will continue to explore the effectiveness of text classification of snippets.

References

[1] R. P. Chhatwal, N. Huber-Fliflet, R. Keeling, J. Zhang, and H. Zhao, "Empirical evaluations of preprocessing parameters' impact on predictive coding's effectiveness," in *Big Data (Big Data), 2016 IEEE International Conference* on, 2016, pp. 1394–1401.

[2] Wei, F., Han, Q., Ye, S., Zhao, H. *"Empirical Study of Deep Learning for Text Classification in Legal Document Review" 2018 IEEE International Big Data Conference.*

[3] N. Huber-Fliflet, J. Zhang, Wei, F., Han, Q., Ye, S., Zhao, H. *"Empirical Comparisons of CNN with Other Learning Algorithms for Text Classification in Legal Document Review" 2019 IEEE International Big Data Conference.*

[4] R. Chhatwal, R. Keeling, P. Gronvall, N. Huber-Fliflet, J. Zhang, and H. Zhao, "CNN Application in Detection of Privileged Documents in Legal Document Review," in *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, Dec. 2020, pp. 1485–1492, doi: 10.1109/BigData50022.2020.9378182.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], Oct. 2018, Accessed: Dec. 01, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805.

[6] Zhao, H, Ye, S., Yang, Jingchao. *"*An Empirical Study on Transfer Learning for Privilege Review*" 2021 IEEE International Big Data Conference.*

[7] Wei, F., Yang J., Mao, Q., Han, Q. *"*An Empirical Comparison of DistilBERT, Longformer and Logistic Regression for Predictive Coding*" 2022 IEEE International Big Data Conference.*

[8] N. Huber-Fliflet, J. Zhang, P. Gronvall, F. Wei. *"*Explainable Text Classification for Legal Document Review in Construction Delay Disputes*" 2022 IEEE International Big Data Conference*

[9] Eugene Yang, Sean MacAvaney, David D. Lewis3, and Ophir Frieder. "Goldilocks: Just-Right Tuning of BERT forTechnology-Assisted Review" arXiv:2105.01044v2 [cs.IR] 19 Jan 2022

[10] V. Sanh et al., DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2019), arXiv:1910.01108.