Leveraging Natural Language Processing and Large Language Models for Assisting Due Diligence in the Legal Domain

Myeongjun Erik Jang[1] **Gábor Stikkel**[2]
[1] Department of Computer Science, University of Oxford, UK
[2] Data Science Lab, Clifford Chance, UK
myeongjun.jang@cs.ox.ac.uk gabor.stikkel@cliffordchance.com

Abstract

Due diligence is a crucial legal process that mitigates potential risks of mergers and acquisitions (M&A). However, despite its prominent importance, there has been a lack of research regarding leveraging NLP techniques for due diligence. In this study, our aim is to explore the most efficient deep-learning model architecture for due diligence in terms of performance and latency, and evaluate the potential of large language models (LLMs) as an efficient due diligence assistant. To our knowledge, this is the first study that employs pre-trained language models (PLMs) and LLMs for the due diligence problem. Our experimental results suggest that methodologies that have demonstrated promising performance in the general domain encounter challenges when applied in due diligence due to the inherent lengthy nature of legal documents. We also ascertain that LLMs can be a useful tool for helping lawyers who perform due diligence.

1 Introduction

Due diligence, one component of mergers and acquisitions (M&A), involves identifying multiple factors that indicate successful outcomes produced by a target organisation (McGrady, 2005). The primary objective of this process is to minimise risks associated with the organisation. Like other legal retrieval tasks, such as contract analysis and cross-jurisdictional analysis, it has been conducted manually by legal professionals. Due diligence is often regarded as a tedious, expensive, and time-consuming job, as the buyer must digest a colossal amount of information within a limited time, often without complete access to relevant information sources (Howson, 2003). However, it is an exceptionally important task, as deficient due diligence can result in significant detrimental outcomes for the buyer [1]. For this reason, there has been a growing demand for automated and precise techniques for due diligence.

The recent remarkable advancements in natural language processing (NLP) field have expanded the potential for developing such techniques. The success of pre-trained language models (PLMs) based on Transformer structure (Vaswani et al., 2017) has led to their application in the legal domain, giving rise to legal-specific PLMs (Chalkidis et al., 2020; Geng et al., 2021; Zheng et al., 2021) and datasets for pre-training (Henderson et al., 2022) and downstream tasks, such as ContractNLI (Koreeda and Manning, 2021) and LexGLUE (Chalkidis et al., 2022). Furthermore, the recent emergence of large language models (LLMs) gained significant attention due to their impressive performance in examinations in legal (Bommarito II and Katz, 2022; Choi et al., 2023) and other professional domains (Terwiesch, 2023; Kung et al., 2023), sparking the possibility of the advent of AI assistants in industrial fields.

However, despite its importance, applying NLP techniques to the due diligence problem has received limited attention. A leading cause would be the lack of publicly available datasets. Due to the nature of M&A, documents for due diligence often contain sensitive information, making it challenging to collect a large-scale dataset. To our knowledge, the KIRA dataset (Roegiest et al., 2018), where the task is designed to detect crucial information in legal contract documents, is currently the only publicly available dataset for due diligence, but it is firmly restricted only to academic usage and obtaining permission to access the dataset requires time and effort. Also, the inherent lengthiness of legal documents poses an additional obstacle. Legal documents often substantially exceed the maximum length that state-of-the-art NLP models can accommodate (Chalkidis et al., 2022), making the models unable to process longer text properly. As a result, most downstream tasks de-

[1] 13 Huge due diligence disasters. [Link]

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pages 155–164

signed to evaluate the performance of legal-specific PLMs have primarily focused on relatively short paragraphs, such as classification (Chalkidis et al., 2022) and question answering (Hendrycks et al., 2021a; Wang et al., 2023).

This paper explores the feasibility of applying modern NLP techniques to the due diligence problem. We first examine the performance of three different architectures on due diligence. Subsequently, we conducted a few-shot experiments on GPT-4 to ascertain whether LLMs could be a useful tool to help the due diligence problem. To the best of our knowledge, this is the first work that leverages PLMs and LLMs for due diligence. Our contributions can be summarised as follows:

- We observe that the hierarchical sentence extraction structure is the most suitable architecture for due diligence and is more practically efficient than the KIRA baseline models.
- We ascertain that legal-specific PLMs do not necessarily outperforms normal PLMs.
- We confirm that LLMs like GPT-4 can be a practical tool to help lawyers conduct due diligence.

2 KIRA Dataset for Due Diligence

Due diligence is a legal process to effectively mitigate the potential risks associated with a company during mergers and acquisitions (M&A). The due diligence problem can be divided into two primary processes: 1) the identification of relevant passages presented in legal documents based on the required information and 2) the utilisation of these passages to predict any potential risks to the acquiring company. Roegiest et al. (2018) collected and released the dataset for the first process exclusively for academic purposes. The dataset contains real-world legal documents across 50 topics, such as "Evidence of Loans" and "Administrative Agent Fees". Each document is transformed into text using Optical Character Recognition (OCR) and other preprocessing techniques. Each sentence within the documents is annotated by KIRA's in-house annotators, including law students, contract lawyers, and in-house senior lawyers. This annotation aims to determine the presence of relevant information in a sentence. The basic statistics of the dataset are presented in Table 1. It is worth highlighting the distinctive characteristics of the dataset, 1) the documents exhibit considerable length, having more than 3K sentences, and 2) the number of relevant sentences is exceedingly scarce. More detailed statistics for each topic are available in Table 7 in the Appendix A. The dataset consists of five folds, where one fold is used for evaluation while the remaining folds are used for training in an alternating fashion. Roegiest et al. (2018) transformed each sentence to human-crafted features and trained a conditional random field (CRF) model that predicts the label of each sentence.

	# of Docs	Doc Length	# of RS	# of Docs w/o RS
Avg	307.7	3308.4	4.8	95.4
Std	94.8	473.5	5.4	69.1

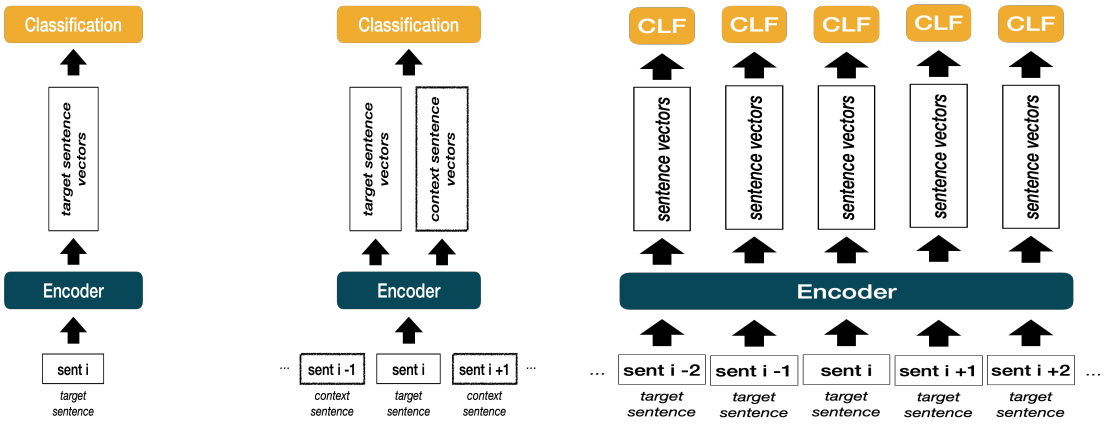
Table 1: Average and standard deviation of basic statistics of KIRA dataset across 50 topics. "RS" denotes relevant sentences, and "Doc Lenght" is the number of sentences in a document.

3 Experiments Design

The KIRA dataset (Roegiest et al., 2018), which serves as the primary dataset in our study, is collected for the first process. It formulates due diligence as a binary sequential classification task, where the relevant sentences in legal documents are labelled by human annotators. The noteworthy characteristic of the KIRA dataset is that the label distribution is highly skewed, where, on average, a document consists of 3300 sentences, but only 4.8 sentences are labelled as "relevant". Here, we explore the due diligence performance of three distinct architectures: 1) single-sentence classification, 2) context-aware sentence classification, and 3) hierarchical sentence extraction. The brief illustrations of these models can be found in Figure 1.

Single-Sentence Classification. This is the simplest-level architecture that considers each sentence independently. The model takes a list of tokens and predicts its label, i.e., "relevant" or "non-relevant". We fine-tune two PLMs: BERT (Devlin et al., 2019) and LegalBERT (Chalkidis et al., 2020).

Context-Aware Sentence Classification. This is an improved version of the single-sentence classification. Following the work of Fang and Koto (2022), the model incorporates the target sentence along with its surrounding sentences to consider a sentence-level context.



(a) Single-sentence Classification (b) Context-aware Classification (c) Hierarchical Sentence Extraction

Figure 1: Illustration of the explored model architectures.

Hierarchical Sentence Extraction Given that the due diligence task aims to extract sentences that deliver relevant information from a document, the most similar NLP downstream task is an extractive summarisation that also selects summary sentences from a document. However, the extensive length of legal documents hinders employing PLM-based extractive summarisation methods, such as BERTSUM (Liu, 2019), because they can only accommodate the limited token length. To address this concern, we adopted a hierarchical structure that effectively handles documents with long lengths (Yang et al., 2020; Chalkidis et al., 2021; Lu et al., 2021). Specifically, the architecture consists of two encoders: a sentence-level encoder that transforms each sentence into fixed-size sentence vectors and a document-level encoder that takes the list of sentence vectors as input and performs a sequential binary classification of whether each sentence contains relevant information.

Training Strategy. We observed that the label distribution is highly skewed (see Table 1 in appendix), which can cause a huge class imbalance issue. We devised a sampling strategy called IMBALANCED SAMPLER to address this concern. The sampler first calculates the probability of an instance with label l_i being chosen in a mini-batch in the following manner:

$$p_i = \frac{N_i}{\sum_{j=1}^{K} N_j},$$

where N_j is the number of training samples labelled l_j. Next, training instances for each mini-batch are sampled using a multinomial distribution, where the probabilities p_i are utilised to determine the sampling with replacement.

On top of the IMBALANCED SAMPLER, we additionally introduced weighted binary cross-entropy loss, as we observed that the class imbalance issue persists. The loss function is defined as follows:

$$\mathcal{L}_{wce} = \sum_{i=1}^{N} \alpha \times y_i \times \log f(x_i) + (1-\alpha) \times (1-y_i) \times \log(1-f(x_i)),$$

where x_i is the i-th instance, f is a model, y_i is the target label for i-th training example, and α is the pre-defined weight.

Training Details. In the single-sentence classification model, both BERT-base and Legal-BERT were trained for three epochs by using AdamW optimiser (Loshchilov and Hutter, 2017) with a learning rate of $5e^{-6}$ and a weight decay rate of $1e^{-2}$. The batch size and maximum number of tokens were set to 32 and 512, respectively. The most important hyperparameter for training is the cross-entropy weight (α). We investigated the optimal α value within a range of {0.7, 0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9} and selected the value that yields the highest validation performance.

The context-aware classification models were fine-tuned with identical training hyperparameter configurations as the single-sentence classification model, apart from using a learning rate of $1e^{-5}$. The optimal α value was determined through exploration within a search space of {0.7, 0.725, 0.75,

Topics	1086			1243			1244			1247			1469		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
BERT-base (*Single*)	.75	.81	.78	-	-	-	.62	.77	.69	-	-	-	-	-	-
Legal-BERT (*Single*)	.79	.85	.82	-	-	-	.38	.89	.54	-	-	-	-	-	-
BERT-base (*Context*)	.67	.87	.75	-	-	-	.50	.61	.55	-	-	-	-	-	-
BERT-base	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
BiLSTM-single-0.5	.90	.85	.88	**.77**	.64	.70	**.81**	.67	.74	**.66**	.67	.67	**.74**	.70	.72
BiLSTM-single-0.9	.89	.89	.89	.74	.71	.73	.77	.73	.75	.62	.74	.68	.71	.75	**.73**
BiLSTM-ensemble-0.5	.90	.86	.88	**.77**	.65	.71	.80	.69	.74	**.66**	.68	.67	.73	.71	.72
BiLSTM-ensemble-0.9	.89	.89	.89	.74	.72	.73	.76	.75	**.76**	.62	.76	.68	.71	.76	**.73**
KIRA-Baseline	**.91**	**.95**	**.93**	.71	**.86**	**.78**	.54	**.91**	.68	.61	**.85**	**.71**	.57	**.89**	.69

Table 2: The performance of different model architectures. "P" and "R" denote precision and recall, respectively. The best performance is highlighted in bold. *Single* and *Context* refer to the single-sentence and context-aware classification models, respectively. Each experiment is repeated five times, and their average is reported. 0.5 and 0.9 denote the cut-off confidence score.

0.775, 0.8, 0.825, 0.85, 0.875, 0.9}.

In hierarchical sentence classification models, we segmented each document into multiple paragraphs to facilitate efficient training. Each paragraph consists of a maximum of k sentences, where the value was set to 16 in our experiments. These paragraphs serve as the basic training units. During the inference phase, predictions were generated for all paragraphs, which were then compared against gold labels to calculate the evaluation metrics.

A model BERT-base as a document-level decoder was trained for 10 epochs with a batch size of 32. AdamW optimiser (Loshchilov and Hutter, 2017) with a learning rate of $1e^{-5}$ and a weight decay rate of $1e^{-2}$ was used for training. The Bi-LSTM document-level decoder models were trained for 30 epochs with a batch size of 32. The learning rate and weight decay rate were set to $1e^{-3}$ and $1e^{-2}$, respectively. Early stopping was applied for both models, whereby the training was halted if the validation performance did not improve for three consecutive epochs. Similar to the preceding experiments, the optimal α value was searched in a search space of {0.7, 0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9}.

The Bi-LSTM document-level decoder models have additional hyperparameters that decide the model's architecture. Below are such hyperparameters and the corresponding search space we investigated to find the optimal values.

- Number of layers (N): 1, 2, 3, 4
- Number of hidden dimension (H): 16, 32, 64, 128, 256
- Dropout rate (Dr): 0.1, 0.2, 0.3, 0.4

Table 3 presents the selected values for each topic. All models were trained using a GeForce GTX TITAN XP GPU. Huggingface transformer package was used for the implementation.

	1086	1243	1244	1247	1469
N	2	1	2	1	1
H	64	16	64	32	64
Dr	0.2	0.1	0.1	0.2	0.1

Table 3: Selected BiLSTM hyperparameters for each topic.

4 Experiments and Results

Single-sentence classification result. We first fine-tuned single-sentence classification models based on BERT and LegalBERT. For the experiment, we chose two topics in the KIRA dataset due to the extensive time and resources needed to conduct experiments on all 50 topics. Specifically, we chose topics 1086 and 1244, where the KIRA-baseline model performed the best and worst, respectively. The experimental results are presented in the second row of Table 2.

The results revealed two important findings. Firstly, both BERT and LegalBERT produced comparable or lower F1 scores than the KIRA baseline, a simple CRF employing human-crafted features. The results indicate that sentence-level sequential information is a crucial factor in the due diligence problem rather than increasing the model complexity. Secondly, LegalBERT did not exhibit a substantial performance advantage over BERT, implying that legal PLMs do not necessarily ensure improved performance in legal-domain downstream tasks. This finding also aligns with the findings of Geng et al. (2021).

Context-aware classification result. Next, we fine-tuned BERT-base with the context-aware architecture (Fang and Koto, 2022) on topics 1086 and 1244. LegalBERT was not included in this experiment because no significant performance difference was observed with BERT-base in single-sentence classification experiments. The performance of the context-aware classification model is presented in the second row of Table 2. Interestingly, even with additional context information, the model performed similarly or worse than the single-sentence classification model. We strongly believe that a leading cause is that accommodating four context sentences is not guaranteed due to the model's maximum length limitation. Our findings suggest the NLP techniques that exhibited favourable performance in general corpora may encounter challenges and limitations when applied to specific industrial fields due to the inherent unique characteristic of the domain.

Hierarchical sentence extraction result. Subsequently, we trained a hierarchical sentence extraction model. On top of the two topics used in preceding experiments, we added three more topics: 1243, 1247, and 1469, where the KIRA-baseline models demonstrated the poorest performance. The other two architectures were not evaluated for these three topics, as they already generated inferior performance than the hierarchical sentence extraction model in topics 1086 and 1244.

When it comes to the sentence-level encoder, we used Sentence-BERT (Reimers and Gurevych, 2019) ALL-MINILM-L6-V2 model [2]. For the document-level encoder, we employed two models: Bi-LSTM and BERT-base. Regarding the Bi-LSTM document-level decoder, we introduced four variations based on the cut-off confidence score (0.5 and 0.9) and single/ensemble methods. The ensemble method made decisions based on majority voting by using the predictions of five models for each test scenario.

The experimental results are presented in the third row of Table 2. Contrary to the common belief that fine-tuned PLMs generally outperform simpler models like Bi-LSTM, BERT-base totally fails to detect relevant sentences. We observed that for all topics, fine-tuned BERT-base predicted all sentences as "non-relevant", a signal indicating the presence of an overfitting issue, which easily occurs in datasets having highly skewed label distribution. The best Bi-LSTM hyperparameters presented in Table 3 also support that the issue of overfitting exists, which shows more layers or hidden dimensions produced worse performance in general. Our findings suggest that increasing the model's scale is not always beneficial when dealing with real-world data.

In topic 1086, the KIRA-baseline model performed the best, but our Bi-LSTM models also produced a decent performance. For the other four topics, while there was no huge difference in terms of the F1 score, our approaches consistently produced substantially higher recall values across all four topics. The high recall model is more efficient than the high precision model from a practical viewpoint in due diligence, where the "relevant" sentences account for an extremely small portion [3] , which can greatly reduce the effort for extensive manual review to detect false negatives. Let us assume that we have 100K sentences and only 100 sentences are relevant. Table 5 shows two extreme cases of high recall but low precision (Case 1) and vice versa (Case 2). For the former, given our awareness that the model attains a high recall rate, it is evident that the majority of relevant sentences are included in the subset of sentences where the model predicts them as "relevant". Therefore, a lawyer can review only 990 sentences (predicted as "relevant") to filter out false positives. However, regarding the latter, the situation is entirely contrasting. While the high precision rate implies that most of the sentences predicted as "relevant" are correctly classified, the low recall rate indicates the presence of numerous false negatives, 90 cases in the example above. Missing 90% of true relevant sentences is very critical, and a lawyer should review nearly 100K sentences to identify false negatives, which would impose an extremely demanding workload. Hence, we can argue that our hierarchical approach is more practically efficient than the KIRA-baseline model in the four topics.

In-context learning with GPT-4. Recently, LLMs has gained huge attention for passing legal examinations, such as the University of Minnesota Law School exam (Choi et al., 2023) and the US bar exam (Bommarito II and Katz, 2022). Hence, we explored how LLMs can be employed to assist with

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] In topic 1244, for example, we can estimate from Table 7 that about 500 sentences are "relevant" while 860K sentences are "not relevant".

PROMPT QUESTION: The definition of Collateral/Transaction Security topic is as follows. 'Lenders will typically require some form of security/collateral to be provided by the borrower or other obligors as a precondition to lending to ensure that if the borrower does not repay the loan or defaults under the credit agreement in any other way, the lenders will have recourse to such security to ensure repayment of the loan. This topic assists in identifying which forms of security/collateral are applicable to a particular transaction.'
Your task is to determine whether given document contains relevant information regarding Collateral/Transaction Security.

Here are samples for this task:
Document: {sample_doc_1}
Answer: {sample_answer_1}

Document: {sample_doc_2}
Answer: {sample_answer_}

Does this document contain relevant information?
Document: {test_doc}
Answer:

Table 4: Prompt used in in-context learning for topic 1243.

		Pred: Case1		Pred: Case2	
		¬R	R	¬R	R
Gold	¬R	99,000	900	99,899	1
	R	10	90	90	10

Table 5: Example confusion matrices for high recall/low precision (Case1) and high precision/low recall (Case2). R and ¬R denote "relevant" and "non-relevant", respectively.

Topics	1243		
	R	P	F1
GPT-4 (2 shots)	.93	.72	.81
GPT-4 (4 shots)	.95	.70	.81
GPT-4 (6 shots)	.95	.72	**.82**
GPT-4 (8 shots)	**.96**	.72	**.82**
KIRA-baseline	.71	**.86**	.78

Table 6: In-context learning performance on topic 1243. The best performance is highlighted in bold.

the due diligence problem. To conduct experiments, we simplified the task into a binary classification that predicts whether a given paragraph contains relevant sentences or not. We tested GPT-4 on topic 1243 by providing a paragraph consisting of 16 sentences. We sampled 100 examples for each experiment, as conducting experiments on the whole dataset is an extensive resource-consuming work. Regarding the prompt design, we first demonstrated the topic definition and task description, followed by two samples. The model was then asked to make a prediction of a new paragraph. The example of the prompt design we used is presented in Table 4.

The experimental results are shown in Table 6. Despite the simplified task transformation, GPT-4 achieved a comparable but lower f1-score than the KIRA-baseline model. However, we observed that providing more few-shot samples can improve the performance, as demonstrated by Hu et al. (2023). Also, GPT-4 exhibited a very high recall rate and a decent level of precision rate, which can greatly reduce lawyers' workload in the due diligence problem, as described above. Implementing a combined system that identifies paragraphs containing relevant sentences through LLMs and then using a high-precision model to detect relevant sentences automatically could further diminish the workload.

5 Related Works

The progress in the field of NLP has been a driving force of the vigorous advancements in legal NLP, leading to a substantial volume of published papers each year since 2017 (Katz et al., 2023). Many legal NLP studies involve predicting judgement decisions (Zhong et al., 2018; Chalkidis et al., 2019; Medvedeva et al., 2020), collecting legal datasets (Zhong et al., 2020; Luz de Araujo et al., 2020; Koreeda and Manning, 2021; Chalkidis et al., 2022) and training legal PLMs (Chalkidis et al., 2020; Geng et al., 2021; Zheng et al., 2021; Xiao et al., 2021; Hendrycks et al., 2021b). However, the application of NLP in due diligence for M&A has not received attention despite its promising importance. Roegiest et al. (2018) collected large corpora to train an automated due diligence model and developed a CRF model to assess the presence of relevant information in each sentence of a legal document. Chitta and Hudek (2019) developed a question answering (QA) system for the due diligence problem, which operates in two phases: 1) identifying *evidence* from a contract that contains the answer to the given question and 2) providing an answer based on the detected evidence. The CRF model developed by Roegiest et al. (2018) is used to find evidence in the first phase. Don-

nelly and Roegiest (2020) employed the same CRF model for named entity recognition (NER) in legal documents, assuming that named entities would exist in sentences containing important information. The CRF model is also utilised by Donnelly and Roegiest (2020) for NER in legal documents. They assumed that named entities in legal documents would exist in sentences containing important information. Therefore, they first used the CRF model to extract candidate sentences, and subsequently trained a named entity detection model using the extracted candidates. This two-step approach demonstrated superior performance in terms of both time and accuracy compared to the state-of-the-art deep-learning NER model of that period (Akbik et al., 2019). The wide adoption of the CRF model suggests that implementing a more accurate relevant sentence extraction model can greatly benefit various legal NLP tasks.

6 Summary and Outlook

Due diligence plays a crucial role in ensuring a successful M&A. Implementing an automated due diligence system will offer significant benefits considering the resources required for due diligence. This paper illuminates the unhighlighted legal NLP topic: the due diligence problem. In this paper, we first explored three neural model architectures: 1) sentence-level classification, 2) context-aware classification, and 3) hierarchical sentence extraction. Subsequently, we examined how GPT-4 can be utilised to assist the due diligence problem. We confirmed that the hierarchical sentence extraction model best suits due diligence and is practically more efficient than the previous approach. Our experimental results indicate that previous traditional approaches should not be underestimated, as they possess valuable merits that can be employed in practical applications to enhance productivity. We also verified LLMs' potential as a useful assistant for lawyers who conduct due diligence.

7 Limitations

Due to the limited computing resources and the enormous size of the KIRA dataset, we focused on five selected topics, which is 10% of the total number of topics the dataset covers. Investigating a broader range of topics could provide more evidence that can support our claim.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Michael Bommarito II and Daniel Martin Katz. 2022. GPT takes the bar exam. *arXiv preprint arXiv:2212.14402*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Radha Chitta and Alexander K Hudek. 2019. A reliable and accurate multiple choice question answering system for due diligence. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 184–188.

Jonathan H. Choi, Kristin E. Hickman, Amy Monahan, and Daniel B. Schwarcz. 2023. ChatGPT goes to law school. *Minnesota Legal Studies Research Paper*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jonathan Donnelly and Adam Roegiest. 2020. The utility of context when extracting entities from legal documents. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2397–2404.

Biaoyan Fang and Fajri Koto. 2022. Context-aware sentence classification in evidence-based medicine. In *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, pages 193–198, Adelaide, Australia. Australasian Language Technology Association.

Saibo Geng, Rémi Lebret, and Karl Aberer. 2021. Legal transformer models may not always help. *arXiv preprint arXiv:2109.06862*.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021a. Cuad: An expert-annotated nlp dataset for legal contract review. In *Advances in Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021b. Cuad: An expert-annotated nlp dataset for legal contract review. In *Advances in Neural Information Processing Systems*.

Peter Howson. 2003. *Due diligence: The critical stage in mergers and acquisitions*. Gower Publishing, Ltd.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in Adam. *ArXiv*, abs/1711.05101.

Jinghui Lu, Maeve Henchion, Ivan Bacher, and Brian Mac Namee. 2021. A sentence-level hierarchical bert model for document classification with limited labelled data. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pages 231–241. Springer.

Pedro Henrique Luz de Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. 2020. VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

Steve McGrady. 2005. Extending due diligence to improve mergers and acquisitions. *Bank accounting and finance*, 18(4):17.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28:237–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Adam Roegiest, Alexander K Hudek, and Anne McNulty. 2018. A dataset and an examination of identifying passages for due diligence. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 465–474.

Christian Terwiesch. 2023. Would Chat GPT get a Wharton MBA? A prediction based on its performance in the operations management course. *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania. Retrieved from: https://mackinstitute. wharton. upenn. edu/wpcontent/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24. pdf [Date accessed: February 6th, 2023]*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. Maud: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1725–1734.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

A Appendix

Topic Number	Topic Name	# of Doc	Doc Length	# of RS	# of Docs w/o RS
1086	Evidence of Loans	78	4595.5	7.6	4
1238	"All-In Yield" Definition	203	4117.1	1.1	118
1239	"Applicable Margin" Definition	318	3552.6	24.2	30
1240	"Base Rate" Definition	407	3429.9	14.0	36
1242	"Cash Equivalents" Definition	318	3552.6	4.7	139
1243	"Collateral"/"Transaction Security" Definition	293	3043.1	4.0	109
1244	"Collateral Documents"/"Security Documents" Definition	253	3416.1	2.0	101
1245	"EBITDA" Definition	367	3425.6	12.6	82
1247	"Coverage Ratio"/"Interest Cover" Definition	318	3552.6	1.8	136
1248	Default Interest - Credit Agreement	290	2938.8	4.0	66
1249	"Defaulting Lender" Definition - Credit Agreement	253	3416.1	3.0	104
1250	"Disqualified Institutions" Definition	233	3999.2	0.6	168
1251	"Currency" Definition	293	3043.1	2.4	32
1252	"Disqualified Stock" Definition	203	3343.7	0.9	137
1253	"Excluded Subsidiary" Definition	516	4025.8	1.9	239
1261	Fundamental Changes Negative Covenant	334	2966.8	6.9	41
1262	Dispositions or Asset Sales Negative Covenant	294	3277.4	13.9	26
1265	Change of Business Negative Covenant	334	2966.8	2.5	48
1267	Burdensome/Restrictive Agreements Negative Covenant	294	3277.4	3.9	164
1272	Accounting Changes Negative Covenant	294	3277.4	1.2	140
1275	Anti-Corruption and Sanctions Covenant	339	3533.3	2.6	168
1300	Financial Statements Affirmative Covenant	374	3546.0	26.4	11
1304	Existence and Conduct of Business Affirmative Covenant	414	3269.4	4.3	35
1308	Books and Records Affirmative Covenant	414	3269.4	4.8	95
1309	Compliance with Laws Affirmative Covenant	414	3269.4	3.0	49
1312	"Change of Control" Definition - Credit Agreement	339	3684.0	5.6	32
1318	"Restricted Subsidiary" Definition	274	3589.1	0.4	211
1319	"Borrowing Base" Definition	452	4155.5	3.7	256
1320	"Excluded Taxes" Definition	224	3562.2	1.8	57
1321	"Indebtedness" Definition	379	3367.4	8.8	43
1439	Breach of Covenants - Event of Default - Credit Agreement	125	2097.9	4.3	8
1440	Cross Default - Event of Default - Credit Agreement	592	3274.4	4.4	37
1443	ERISA Events - Event of Default - Credit Agreement	376	3339.2	1.8	153
1444	Change of Control - Credit Agreement	252	2795.5	10.2	26
1460	"Specified Representations" Definition	196	3348.9	1.2	73
1462	"Change in Law" Definition	359	4373.4	1.8	68
1468	Commitment Fees - Credit Agreement	232	3106.2	4.4	68
1469	Facility Fee	415	3022.5	3.7	238
1474	Administrative Agent Fees	232	3106.2	1.5	72
1475	Several Liability	232	3106.2	2.6	69
1489	Financial Statements Representation - Credit Agreement	244	2828.3	3.9	38
1498	Environmental Representation - Credit Agreement	244	2828.3	4.1	84
1500	Full Disclosure Representation - Credit Agreement	244	2828.3	3.5	42
1509	Assignment Transfer Fees - Credit Agreement	367	2634.7	0.8	153
1512	Eligible Assignees	367	2634.7	1.0	181
1520	"Approved Fund"/"Related Fund" Definition	375	2685.6	0.5	200
1524	Costs and Expenses	172	2505.9	7.8	10
1551	"Excess Availability" Definition	317	3380.8	0.8	222
1601	Equity Cure Rights	201	3441.7	7.5	31
1611	"FATCA" Definition	327	3616.5	1.1	118

Table 7: Detailed statistics of KIRA dataset for each topic. “RS” denotes relevant sentences, and “Doc Lenght” is the number of sentences in a document. “Doc Length” and “# of RS” is the average value.