Artificial Intelligence and Law
https://doi.org/10.1007/s10506-025-09484-4

ORIGINAL RESEARCH

Cost–benefit analysis of deploying shallow, deep learning and generative models for legal text classification

Eoin O'Connell[3] · William Duffy[1] · Niall McCarroll[1] · Katie Sloan[3] · Kevin Curran[1] · Eugene McNamee[2] · Angela Clist[3] · Andrew Brammer[3]

Received: 26 July 2024 / Accepted: 2 September 2025

Abstract

Recent advances in Generative Language Models (GLMs) have renewed focus on promising results in zero-shot text classification. However, their off-the-shelf performance on unfamiliar and domain specific tasks remains uncertain. In this legal clause classification task we evaluate a plug-and-play zero-shot prompting strategy for OpenAI's GPT-4 GLM on a contract clause dataset. We introduce the new CUAD-SL dataset that has been refactored as a single label classification problem as a fairer and more robust legal classification benchmark. In a comparative study, we show that fine-tuning on legal domain data adapts smaller, less complex models to the task at hand, with significant classification accuracy improvement of up to 20.6%, with a best overall performance of 87.8% for the DeBERTa Transformer model compared to GPT-4's 67.2% performance. This study also takes the novel approach of assessing the business feasibility of deploying each of these machine learning models through a detailed cost–benefit analysis that measures the trade-off between performance metrics and low and high usage running costs.

Keywords Generative language models · OpenAI · Zero-shot learning · Legal clause classification · Stochastic gradient descent (SGD) · BERT · RoBERTa · DeBERTa · GPT-4 · Curie-FT · Ada-FT

✉ Niall McCarroll
n.mccarroll@ulster.ac.uk

[1] School of Computing, Engineering and Intelligent Systems, Ulster University, Derry, Northern Ireland

[2] School of Law, Ulster University, Belfast, Northern Ireland

[3] Donegall Quay, A&O Shearman, Belfast, Northern Ireland

Published online: 11 November 2025

1 Introduction

Text classification is an important supervised machine learning task for assigning predefined labels to different documents, which helps to structure, organise and categorise them. Due to privacy safeguarding and cost issues, there is a lack of publicly available annotated legal data, which has hindered the development of robust systems for classifying legal texts. This low data scenario has restricted the potential of deep learning models and delayed the legal domain's big data revolution relative to other fields of study (Costa et al. 2023; Yan et al. 2019).

Large language models have revolutionized the field of natural language processing, achieving significant performance gains across the NLP task landscape (Brown et al. 2020; Devlin et al. 2019; Gera et al. 2022; Raffel et al. 2020). State-of-the-art generative language models and large pre-trained Transformers have achieved significant gains across many complex NLP tasks, such as classification, and have recently surpassed human performance on tasks such as SuperGLUE (Wang et al. 2019) and SQuAD 2.0 (Rajpurkar et al. 2018).

Recent advances in large pre-trained GLMs have highlighted their promising performances as zero-shot and few-shot classifiers with many researchers speculating on an ambitious vision of a new paradigm where large general purpose zero-shot models can tackle multiple tasks without requiring labelled data. While holding great promise in eliminating the burden of collecting domain and task-specific labelled data, others argue that these models exhibit mediocre plug-and-play performance in unfamiliar tasks compared to models trained conventionally within a supervised learning paradigm (Gera et al. 2022).

There also remains an open question as to whether large transformer models can transfer to highly specialised domains such as law. These deep learning models require thousands of costly annotations and the need to collect labelled data for each target task presents an obstacle, restricting the use of language models in practice, at scale (Bayer et al. 2022; Hendrycks et al. 2021). If machine learning models are only as powerful as the quality of the data that feeds them, then it is plausible that smaller and more shallow models, purpose-trained on legal domain-specific data – which is more homogeneous than natural language – may exhibit competitive performance compared to multi-purpose large language variants.

In this paper, we evaluate the applicability of Generative Language Models (GLMs) for zero-shot legal clause classification. We compare and contrast this zero-shot performance against that of a task-specific trained shallow SGD model and that of larger pre-trained BERT-based transformer and generative OpenAI GPT-3 models that have also been fine-tuned with task-specific legal training data.

As well as contributing to the emerging research area of applying generative and transformer models to unfamiliar and specialised legal domains, another significant contribution of this study is the repurposing of the Contract Understanding Atticus Dataset (CUAD) (Hendrycks et al. 2021) from a multiple label question-answering clause dataset to a more robust and refined single label classification problem, providing a novel benchmark for the testing of shallow and deep learning classifiers.

Finally, machine learning research is only valuable to real-world enterprise when we fully evaluate the value of these models in the context of the trade-off between their performance and the cost to host and deploy these models (Collier et al. 2019). We present a novel and detailed cost–benefit analysis that re-evaluates the performances of each shallow and deep learning model across low and high usage regimes in an effort to re-assess performance metrics within the framework of business value, impact and feasibility.

As the majority of research and development has focused on the exploration of generic text domains, there have been very few programmes that concentrate on niche NLP tasks such as legal text classification. The continuing lack of sufficient amounts of legal training data has placed a new emphasis on the plug-and-play potential of large pre-trained GLMs to perform multiple NLP tasks without requiring labelled data. Within this context, the research objective of this paper is as follows: (1) To propose a Zero-Shot legal clause classification strategy, requiring only natural language prompting and a list of potential class names, for OpenAI's groundbreaking GPT-4 model. (2) To compare and contrast the performance of this Zero-Shot strategy against that of a range smaller neural network models that have been exposed to and trained on legal-specific data. (3) To assess each of these models in a practical business context by performing a unique cost–benefit analysis that appraises the performance of each model within high and low usage scenarios to evaluate and understand the delivery of business value.

The paper is organised as follows: The Sect. 2 literature review compares shallow and deep learning models for legal text classification before assessing emerging trends in GLMs and zero and few-shot learning. After introducing related work, the Methodology section records the extensive work that went in to restructuring the CUAD (Hendrycks et al. 2021) legal clause dataset into the new CUAD-SL (single label) dataset consisting of 23 different clause categories for benchmarking. The Methodology section also details the different shallow, deep and generative models that were used in this evaluation before providing an overview of the cost–benefit analysis that was employed to assess the performance of each model in the context of costs to train, host and deploy these solutions (Sect. 3). Section 4 presents the experimental results documenting the performance of each of the shallow, deep and generative models both in terms of an accuracy performance metric and then through an extensive real-world business analysis of costs versus performance. Section 5 discusses the implications of the results in detail and Sect. 6 concludes with an assessment of how zero-shot learning could be boosted with additional support strategies and whether it is best placed as a data augmentation technique for synthesising additional training data for low resource NLP scenarios. Appendix 1 contains full details, justifications and a glossary of terms for the mapping of the original CUAD clause labels to the new single label CUAD-SL set up.

2 Related research

Text classification is a fundamental research topic in NLP, Machine Learning and text mining and it has important real-world applications within eDiscovery and wider legal document processing (Chen et al. 2022; Graham et al. 2023; Tavor et al. 2020). For various tasks that involve classifying text, rule-based methods and simple statistical approaches like Logistic Regression, Support Vector Machines (SVMs), Decision Trees and Naive Bayes, have often been seen as robust and reliable baseline techniques (Chiticariu et al. 2013; Joachims 1998; McCallum and Nigam 1998; Nallapati and Manning 2008; Wilcox and Hripcsak 1999; Yang and Liu 1999). They can be used on moderately difficult tasks such as classifying the political ideology of a judge based on their judicial opinion texts (Hausladen et al. 2020) and with suitable feature engineering, often achieve state-of-the-art performance and also have the potential to scale to very large corpora (Wang and Manning 2012).

Traditionally, most researchers used simple bag-of-word unigrams, bigrams or n-grams as feature inputs (Dumais et al. 1998). The main disadvantage of these methods is that they disregard contextual information and sequential text structure (Li et al. 2020), which may limit their generalisability to larger output feature spaces (Joulin et al. 2017; Nallapati and Manning 2008).

More recent approaches to NLP have centred around the use of deep learning architectures to achieve Language Modelling (LM), which has revolutionised the field of NLP, resulting in significant leaps in performance across the entire NLP task landscape (Brown et al. 2020; Devlin et al. 2019; Gera et al. 2022; Raffel et al. 2020; Meng et al., 2020). However, Clavié and Alphonsus (2021) argue that there is too much focus on comparing performance within deep learning models rather than comparing them with well-optimised, shallow baselines. The authors experimentally confirm that SVM classifiers reach competitive performance with pre-trained BERT-based models on multiple legal text classification tasks in the LexGLUE (Chalkidis et al. 2022) benchmark. They also note that the relative performance improvement between BERT-based and SVM models is noticeably smaller within the legal-specific domain than on general text classification tasks, even when legal-domain specifically trained models such as Legal-BERT (Chalkidis et al. 2020) and CaseLaw-BERT (Zheng et al. 2021) are used (Clavié and Alphonsus 2021).

More recently, Chen et al. (2022) demonstrated that Random Forests using Domain Concept features outperformed a BERT-plus-LSTM model in the task of categorising 30,000 U.S. case documents into 50 different groups (Chen et al. 2022). Deep learning models also require much greater compute resources as well as larger, carefully curated datasets to achieve optimum performance (Yan et al. 2019). This has hindered their adoption in applied practice areas such as law where client privacy concerns and the lack of publicly available datasets has limited access to training resources.

Generative Language Models (GLMs) have become the tool of choice for any and every task with state-of-the-art benchmark performances being documented in a number of NLP studies (Laskar et al. 2023). Existing studies have explored zero and few-shot learning with these models for tasks such as machine translation (Gu et al. 2018), instruction following (Branavan et al. 2009; Chen and Mooney 2011), and structured query generation (Huang et al. 2018). However, there is an open question as to whether these extremely large models are the right choice for domain-specific tasks and whether the size of these models and their training datasets correspond to quality and performance (Lu et al. 2022).

Zero-shot text classification is a machine learning technique that enables models to classify inputs from previously unseen classes, without having seen any specific training data on those classes. Radford et al. (2019) demonstrate that large language models can be useful for zero shot classification tasks and that larger models generally perform better at these tasks. They do highlight however that these models require extensive prompt engineering to guide the model responses (Radford et al. 2019; Schopf et al. 2022).

Wei et al. (2022) further argue that the quality of data is more important than the quantity and that the problem with large language models may relate to their training data. GLMs are pre-trained on large amounts of general-domain web-crawled text data such as news articles and Wikipedia. As a result, these models may not perform well in specific domains, such as biomedicine or law compared to smaller, simpler models that are fine-tuned on high-quality, relevant data (Wei et al. 2022). Even then, there is little consensus on the impact of pre-training or fine-tuning on domain-specific data. Lu et al. (2022) found that a T5 model pre-trained on clinical text outperforms the T5 base model in clinical domain-specific tasks and compares favourably with its close baselines (Lu et al. 2022). However, Moradi et al. (2021) demonstrate that even when fine-tuned on biomedical data, GPT-3 models struggle to identify relationships, answer questions and classify text on a par with well-tuned models that are orders of magnitude smaller (Moradi et al. 2021).

As language models become larger and more capable, they also become more costly and inefficient (Ding et al. 2024). Training requirements from a hardware and monetary perspective will become prohibitively expensive if the trend continues. These sophisticated models benchmark well within academic research papers but might not pay off when we appraise their performance within the context of a business use case. This study will take a novel approach of evaluating the predictive performance of each shallow, deep and generative model from an investment perspective where the costs of training, hosting and deploying these models are measured against the benefit of their predictive performance in low and high usage scenarios.

3 Methodology

3.1 Dataset

The dataset used for all experiments is the Contract Understanding Atticus Dataset (CUAD) (Hendrycks et al. 2021). This commercial contract dataset is curated and maintained by legal experts at The Atticus Project (Hendrycks et al. 2021; The Atticus Project, n.d.) in an effort to support NLP research and development in legal contract review. The original purpose of the dataset was to label sections and clauses within contracts which are important for human review. Example labels include, Effective Date, Renewal Term and Governing Law. In all, the original dataset of 510 contracts contains 41 label categories and over 13,000 annotations.

The CUAD dataset in its existing format has a number of features that were inconsistent with the current experimental set up for single-label clause classification. Following the SQUAD v2.0 (Rajpurkar et al. 2016) labelling method, CUAD was labelled by dozens of law student annotators using a combination of short and long spans, from full paragraphs to sub-strings, leading to a significant amount of overlap and inconsistency with multiple spans of text of varying lengths having multiple labels.

Therefore, the first task in this experimental set-up was to employ a more consistent approach to segmenting the different contracts within the dataset. This study uses a version of the GraphSeg algorithm for text segmentation (Glavas et al. 2016). GraphSeg exploits the semantics of text through word embeddings and a cosine measure of semantic relatedness of text to construct a semantic relatedness graph of text chunks – in this case, the extracted sentences within the legal contracts. The next stage was to assign each segment a single label. Since several of the labels tended to co-occur, we selected the dominant label as the one with greatest coverage within a segment. This is simply measured in terms of the greatest number of characters contained within a span. When there is more than one label with equal greatest coverage, all labels are provisionally assigned.

Through the assistance of a legal domain expert and for the sake of full transparency, the authors map the original multi-label CUAD dataset labels to the new single label set-up, providing evidence and clause definitions to justify each decision. Appendix 1 contains full details, justifications and a glossary of terms for the mapping of the original CUAD clause labels to the new single label set up. Table 1 provides a high-level overview of original label to new label mappings. The new modified CUAD dataset will be known as *CUAD-SL (Single Label).*

The majority (around 90%) of our derived segments in the dataset are unlabelled. This imbalance severely skews the class distribution and has the potential to introduce significant misclassification costs, particularly in the case of the generative language models. It was on balance better to exclude them from the benchmark since we are not trying to train a perfect classifier but rather compare a number of different kinds of models on an even footing. The training:validation data ratio was 7:3, stratified by class. The number of words in each segment was comparable between the training and validation sets with a median [IQR] of 10 [4, 28] and 11 [4, 30], respectively.

3.2 Models

Within legal text classification research, there has been very little research to compare the performance of state-of-the-art generative language models, well-established transformer models, and solid baselines like Stochastic Gradient Descent (SGD) classifiers which tend to perform well on simple text classification tasks (Diab 2019). We compare the performance of a TF-IDF + SGD pipeline, BERT, RoBERTa and DeBERTa transformer models, OpenAI's fine-tuned Ada and Curie GPT-3 models, as well as zero-shot classification with GPT-4.

In its current release, GPT-4 is not available for fine tuning and foundational Ada and Curie models do not have the ability to follow instructions. Fine-tuning the smaller GPT models in theory should perform better than zero-shot prompting particularly with a relatively large number of classes. The hyperparameters for each model were selected based on best practices established in the literature.

Table 1 Mapping and counts of the original 41 CUAD labels to the new 23 single label set-up in the new CUAD-SL dataset

Original CUAD Labels	New Label	Train (n)	Test (n)
Anti-assignment Non-Transferable License	Transfer Restrictions	733	315
Cap on Liability Uncapped Liability	Liability Limit	696	299
Document Name Parties Agreement Date Effective Date Expiration Date Renewal Term	Contract Details	1878	805
Insurance Liquidated Damages	Insurance and Liquidated Damages	596	256
IP Ownership Assignment Joint IP Ownership Source Code Escrow	IP Ownership	412	177
Notice Period to Terminate Renewal Termination For Convenience	Termination Rights	284	123
Non-Disparagement Non-compete Exclusivity No-Solicit of Customers No-Solicit of Employees	Exclusivity	641	275
Irrevocable Or Perpetual Licence Affiliate Licence-Licensee Affiliate Licence-Licensor	Licence Terms	168	72
Licence Grant Unlimited/All-You-Can-Eat-License	Licence Scope	319	137
Governing Law	Governing Law	339	146
Most Favoured Nation	Most Favoured Nation	30	14
Competitive Restriction Exception	Competitive Restriction Exception	86	37
ROFR/ROFO/ROFN	ROFR/ROFO/ROFN	307	132
Change of Control	Change of Control	212	91
Revenue/Profit Sharing	Revenue/Profit Sharing	413	177
Price Restrictions	Price Restrictions	21	9
Minimum Commitment	Minimum Commitment	376	162
Volume Restriction	Volume Restriction	145	63
Post-Termination Services	Post-Termination Services	366	158
Audit Rights	Audit Rights	535	230
Warranty Duration	Warranty Duration	159	69
Covenant Not to Sue	Covenant Not to Sue	135	59
Third Party Beneficiary	Third Party Beneficiary	30	13

3.3 OpenAI – GPT-4 8 K context

GPT-4 has better language capabilities than any previous GPT variant, allowing it to adapt more readily to a wider range of tasks. It is optimised for chat but chat is just a broader type of traditional completion tasks. Semi-structured prompts were sent to the GPT-4 API and the completions then post-processed to produce clean labels. The prompt instructs the model to classify the input text using the list of 23 possible labels from the CUAD-SL dataset. The prompt structure is detailed in Appendix 2.

3.4 OpenAI – Curie & Ada GPT-3 fine tune parameters

We employed the standard, widely adopted fine-tuning procedures as implemented in the OpenAI API library. Relevant implementation and specification detail on the Curie and Ada models is available from the OpenAI site (OpenAI, n.d.). For both fine-tunes, the default/inferred parameters from the Azure ML OpenAI Service (Microsoft 2023) were used and are defined as follows: Epochs: 4; Batch size: 8; Learning Rate Multiplier: 0.2; Prompt loss weight: 0.1.

3.5 TF-IDF + SGD training

The TF-IDF + SGD classifier represents a well-established and robust baseline for text classification tasks, including those in the legal domain. This approach leverages n-gram features to capture local word patterns and uses SGD for efficient optimisation, particularly suitable for high-dimensional sparse data typical of text corpora. Prior research has demonstrated that such shallow models, when properly tuned, can achieve competitive performance on legal text classification tasks, sometimes rivalling more complex deep learning models (Clavié and Alphonsus 2021; Chen et al. 2022). The choice of this model was motivated by its interpretability, computational efficiency, and its proven effectiveness as a baseline in both general and legal NLP benchmarks.

The TF-IDF + SGD pipeline is defined as follows: 1–3 ngram count vectoriser with removal of English stop words; TF-IDF transformer; SGD classifier with alpha = 0.0000825, modified_huber loss, 1200 max iterations, random state of 42, tolerance set to None, and balanced class weights.

3.6 BERT variants—BERT, RoBERTa & DeBERTa parameters

The BERT variants used in this study were the BERT base model (cased), the RoBERTa large model and the DeBERTa large model. The selection of BERT and its derivatives was driven by their status as state-of-the-art transformer-based models for a wide range of NLP tasks, including text classification. The use of these three variants allows for a comprehensive comparison across different levels of model sophistication and pre-training strategies. This is particularly relevant in the legal domain, where recent studies have shown that the performance gap between shallow models and deep learning models is often smaller than in general NLP tasks, but that advanced transformer models can still offer meaningful improvements, especially when fine-tuned on domain-specific data (Chalkidis et al. 2020).

BERT-Based Models: For the BERT-based models, we followed the default fine-tuning pipeline as described in the original BERT paper (Devlin et al. 2019) and as implemented in the Hugging Face Transformers library. This approach involves adding a task-specific classification head to the pre-trained BERT model and fine-tuning all parameters on the downstream classification task using supervised learning. Full implementation, specification and benchmark performance details of these models are available from the Huggingface and (Huggingface, n.d.) Azure ML websites (Price et al. 2021). Additional parameters for all three BERT-based models were as follows. The PyTorch framework is used with the transformers Huggingface library for training of the model. The training parameters are 5 for the number of epochs, random seed of 42, 1000 for batch size, 20 for the warmup steps, AdamW_hf optimization, 0.00002 for the learning rate, a weight decay of 0 and 500 evaluation steps.

3.7 Cost–benefit analysis

One of the main goals of this study was to assess the performance of the different shallow and deep learning strategies within a business context by performing a cost–benefit analysis in both a low and high usage scenario. The analysis assumes that all models will be deployed within the Azure Cloud Computing platform where the main costs will be hosting and compute. The additional costs of annotation of training and testing data by junior legal professionals was also included in the analysis. Full analysis of Cost-Per-Day versus model accuracy across both low and high usage scenarios is detailed in the Results section.

4 Results

This research study involves a large multi-class text classification problem involving 23 different clause classes and an imbalanced dataset ranging from a count of 513 instances in the most common ‘Contract Details’ clause to 7 instances in the lowest frequency ‘Price Restrictions’ clause. It was therefore decided that the most useful comparison to evaluate the performance of the different models, and to further consider cost–benefit analysis within a business context, was to calculate a weighted average across the imbalanced data. Weighted accuracy is the overall accuracy that can be expected across a large enough sample with a similar distribution of classes to the test dataset, rather than the probability of correctly classifying a randomly selected clause. Table 2 displays the accuracy results for the shallow SGD model, the deep learning BERT-based transformer variations and the OpenAI Generative models on the CUAD-SL dataset.

Table 2 Overall weighted accuracy performance and standard deviation of shallow, deep, and generative models on the refactored CUAD-SL dataset

Model	Weighted Accuracy (%)	Standard Deviation of accuracy across classes (%)
GPT-4	67.2	24.1
Ada-FT	83.8	17.0
Curie-FT	81.7	15.8
TF-IDF + SGD	83.7	29.0
BERT	78.9	33.6
RoBERTa	86.4	16.2
DeBERTa	87.8	13.4

For more detailed results exploring recall and precision for all labels, see Table 4—Appendix 3

Zero-shot GPT-4 significantly underperforms compared to the other models achieving an accuracy of 67.2% compared to BERT which achieves 78.9% accuracy, with the remaining models achieving accuracy performances of over 80% ranging up to the top performing DeBERTa at 87.8% accuracy. It should be noted however, that an accuracy of 67.2% from a GPT-4 model that is employing zero-shot learning in a specialised and complex legal domain clause classification task, represents an impressive baseline performance. The larger standard deviation for GPT-4 results indicates a lot of variance in performance across the clauses and that it was not worse across all categories as its weighted accuracy score would suggest. Figure 1 provides a more detailed, clause-by-clause breakdown analysis of the performance of each model. GPT-4 achieves 100% accuracy on Governing Law clauses and ~94% accuracy on Audit Rights and Termination Rights clauses and ~92% accuracy on Third Party Beneficiary Clauses (of which there are only 12 instances in the CUAD dataset). GPT-4 also significantly outperforms the other models on classification of the 'Price Restrictions' clauses. GPT-4 scores 85.7% accuracy on these clauses with its nearest competitors being Curie-FT and DeBERTa that both achieved only 57.1% accuracy. OpenAI have not released technical details with regards to the training of GPT-4 but these high accuracy scores would indicate that the model has been exposed to these types of clauses or contracts in its pre-training phase.

From Fig. 1 we can also see a general decline in performance across the final six clauses on the chart. All models underperform on the Licence Scope, Volume Restriction, Licence Terms clauses with the worst overall average performance for all models being 29% on the Competitive Restriction Exception clauses. The decline in performance is likely attributable to smaller amounts of training data being available for these clause categories. Smaller datasets typically contain less details and results in the classification models failing to generalise patterns in the training data (Prusa et

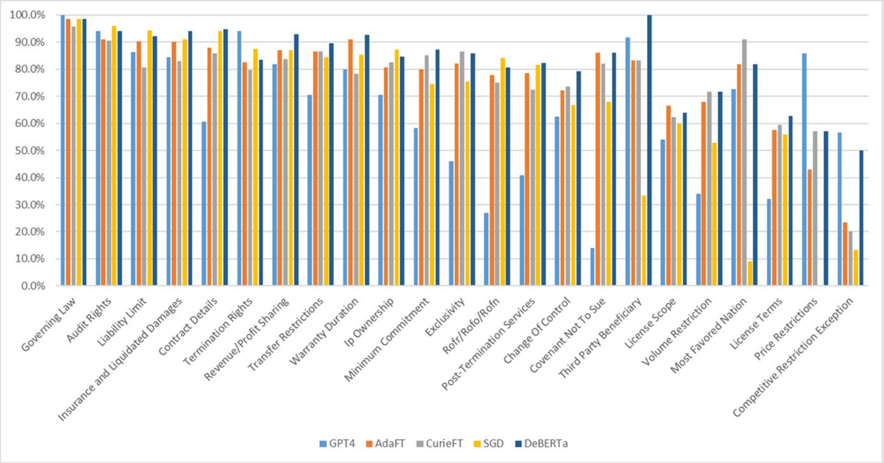

Fig. 1 Chart displaying the Clause-by-Clause classification accuracy of shallow SGD, Deep Learning DeBERTa Transformer and the OpenAI GPT4, Ada and Curie models on the refactored single label CUAD-SL dataset of legal contract clauses. Diagram only showing results for DeBERTa as this was the top performer from the BERT-based trials

al. 2016). It is possible that the lower representation for these type of clauses in the training data is due to them being uncommon and less frequently used compared to the standard boilerplate clauses.

It should be noted that on the two clause categories that all models collectively performed worst on—Price Restrictions with an average overall model performance of 40.8% and Competitive Restriction Exception with an average overall model performance of only 29%—GPT-4 achieves the best overall performance compared to the other models scoring 85.7% on Price Restrictions and 56.7% on Competitive Restriction Exception.

The large standard deviation across the different clauses for the BERT and SGD models contrasts the extremes of classification performance which sees both models achieve 90%+on the top six clause categories and as low as 0% on the bottom six categories. BERT's performance is notable in that it achieves 0% classification on four of the clause categories—Third Party Beneficiary, Most Favoured Nation, Price Restrictions and Competitive Restriction Exception.

To extend the analysis to other models in the OpenAI GPT family, the Curie GPT-3 and Ada GPT-3 foundational models were fine-tuned on the CUAD text-label training pairs. Fine-tuning enables the base models to train on more examples that can fit in a standard prompt and was expected to yield higher quality results than zero-shot prompting. Ada-FT (83.8% accuracy) and Curie-FT (81.7% accuracy) significantly outperform zero-shot GPT-4 (67.2% accuracy). Being a larger model with more advanced language understanding capabilities, Curie-FT would be expected to outperform Ada-FT.

Accuracy performance increased as expected as we move through testing of the different enhanced BERT models. The baseline BERT model achieved 78.9% accuracy, the updated RoBERTa model scored 86.4% accuracy and the state-of-the-art DeBERTa model achieved the top performance of all models tested achieving an accuracy of 87.8%.

Compared to the larger and more complex deep learning transformer and generative models, the shallow SGD is competitive at 83.7%, outperforming BERT (78.9%), Curie-FT (81.7%), GPT-4 (67.2%) and achieving a similar performance to Ada-FT (83.8%) (Table 2).

4.1 Cost–benefit analysis

Table 3 details a comprehensive breakdown of the costs required to run and query each of the shallow, deep and generative models that were evaluated in this study. We assess a low usage scenario with an inference rate of~1 inference/min over a working year of 1,650 h (25m tokens in total) and a high usage scenario with an inference rate of~120 inferences/min over a working year of 1,650 h (3B tokens in total) for the Open-AI generative models. We apply a 50% token mark-up for the GPT-4 model as we have to provide additional context instructions within the prompt to explain the input clause text and choice of labels for the zero-shot learning trial. Non Open-AI models can be deployed on dedicated compute resources and so there is no per-token billing.

Table 3 Cost breakdown analysis for low and high usage regimes for training and test data preparation and the hosting and deployment of the different shallow, deep learning and OpenAI Generative models on the Azure Cloud Computing platform

	Details	GPT4-8 k	AdaFT	CurieFT	SGD	BERT	RoBERTa	DeBERTa
Training Data	7104 text-label pairs @ £0.50 each	-	£ 3,552	£ 3,552	£ 3,552	£ 3,552	£ 3,552	£ 3,552
Test Data	3045 text-label pairs @ £0.50 each	£ 1,523	£ 1,523	£ 1,523	£ 1,523	£ 1,523	£ 1,523	£ 1,523
Training Compute	Standard_NC24s_v3 @ £2.25/hr (low priority) for non-OpenAI models AdaFT is per token + training time	-	£ 14	£ 26	£ 1	£ 3	£ 6	£ 17
Testing Compute	Standard_NC24s_v3 @ £2.25/hr (low priority) for non-OpenAI models OpenAI are token-based (+hosting for FT) 50% token markup on GPT-4 for additional prompt context	£ 27	£ 0	£ 1	£ 0	£ 2	£ 3	£ 5
Low usage								
Hosting (1 year)	1 dedicated Standard_NC6 @ £1/hr for non-OpenAI models GPT-4 base model has no separate hosting costs	-	£ 355	£ 1,706	£ 8,766	£ 8,766	£ 8,766	£ 8,766
Inference costs	25 m tokens (~1 inference/min over a working year of 1650 h) Non-OpenAI models have no per-token billing 50% token markup on GPT-4 for additional prompt context	£ 938	£ 8	£ 41	-	-	-	-
Total cost (1 year)	Assuming gathering new data is an annual cost to prevent stale models	£ 2,487	£ 5,453	£ 6,848	£13,841	£13,845	£13,850	£13,862
Cost per day		£ 7	£ 15	£ 19	£ 38	£ 38	£ 38	£ 38
High usage								
Hosting (1 year)	3 dedicated Standard_NC6 @ £1/hr for non-OpenAI models GPT-4 base model has no separate hosting costs	-	£ 355	£ 1,706	£26,298	£26,298	£26,298	£26,298
Inference costs	3B tokens (~120 inferences/min over a working year of 1650 h) Non-OpenAI models have no per-token billing 50% token markup on GPT-4 for additional prompt context	£112,500	£ 975	£ 4,866	-	-	-	-
Total cost (1 year)	Assuming gathering new data is an annual cost to prevent stale models	£115,958	£ 7,162	£15,186	£48,943	£48,949	£48,955	£48,968
Cost per day		£ 317	£ 20	£ 42	£ 134	£ 134	£ 134	£ 134

Low usage hosting of the non-OpenAI models is calculated on a single Azure dedicated Standard_NC6 Virtual machine costing £1/hr. High usage hosting assumes the requirement for three Standard_NC6 Virtual machines costing £1/hr. The GPT-4 base model has no separate hosting costs.

Annotation of training and test data was estimated at £0.50 per text-label pair based on the average wage of a junior legal professional working at a rate of £15/hour and an estimated labelling time of 30 s per label. There are no training data costs for the OpenAI GPT-4 model as it is being evaluated as a zero-shot learner. Training and testing compute for the OpenAI Fine-tune models and non-OpenAI models is calculated using Azure Standard_NC24s_v3 GPU virtual machines at a low priority rate of £2.25 per hour. Ada-FT and Curie-FT have additional per token costs on top of training and testing compute.

In the low usage scenario, the OpenAI models are much more cost efficient to run with the most expensive Curie-FT model (£19 Per Day) being half the rate of each of the non-OpenAI models (£38 Per Day). GPT-4 stands out at the lowest rate of £7 per day for the low usage inference rate of ~ 1 inference/minute (25 million tokens per year). GPT-4 has no separate hosting costs and simply charges a pay-as-you-go consumption charge per 1,000 tokens used. While this pricing structure is advantageous for GPT-4 is a low usage scenario, we start to see a significant increase in daily rate when we move to the high usage scenario of 3 billion tokens per year. In this scenario, GPT-4 jumps to a daily rate of £317 per day. In a high usage scenario, the requirement for three dedicated Azure Standard_NC6 Virtual Machines for hosting of non-OpenAI models, results in the daily cost of the shallow SGD and the deep learning BERT-based models increasing from £38 per day to £134 per day. The increase in daily rate is much more modest for the Ada-FT and Curie-FT models when the move from a low to a high usage scenario is considered. Overall, Ada-FT remains the most cost-efficient model in terms of hosting and usage, ranging from a Cost Per Day of £15 under a low usage scenario to a small increase of £20 per day under a high usage scenario.

The chart in Fig. 2 combines the low and high usage costings with model accuracy performances on the CUAD-SL Legal dataset trials to provide a useful comparison of the trade-off between cost efficiency and performance metrics for machine learning in a business setting.

Figure 2 provides a very useful overview of the cost–benefit trade-off across the shallow and deep learning transformer and generative models. While GPT-4's zero-shot performance was impressive on a previously unseen dataset, an accuracy of 67.2% rules it out as useful classifier, particularly when we also consider the prohibitive daily pricing under a high usage scenario. It's simply not worth the time, money and resources for this kind of classification task when we consider the performance of the smaller models.

For business scenarios where usage will be modest and higher accuracy is essential, then DeBERTa at the low usage cost of £38 per day and accuracy of 87.8% is a good option. However, if usage increases, this option can get significantly more expensive at £134 per day, which may be beyond the means of smaller firms and enterprises that run on tighter budgets. If you are a business that can slightly more flexible on accuracy, Ada-FT offers a good cost–benefit trade-off across all of the

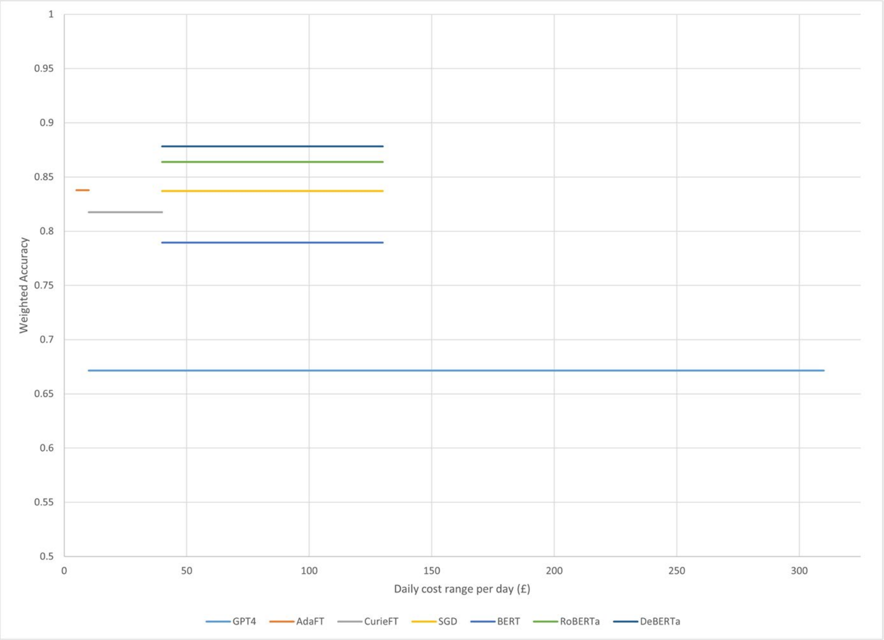

Fig. 2 Comparison of model accuracy performance and low to high daily cost range per day for each of the shallow, BERT-based transformer and OpenAI Generative models that were evaluated on the CUAD legal clause classification study

models evaluated. Performing at an accuracy of 83.8% and a cost per day that ranges from £15 at low usage to £20 per day at the high usage rate, a user can be confident that the model is always going to remain within a manageable cost range regardless of the increase in demands for the service.

5 Discussion

There are three novel phases of investigation in this study that address important gaps within legal automation research. The first phase was the repurposing and creation of the new single label CUAD dataset for fair and robust benchmarking within the legal technology domain. We then performed the unique comparison of state-of-the-art generative language models, deep learning BERT-based transformers and a shallow SGD model on a series of fine-tuned and zero-shot learning strategies for legal clause classification. Finally, this study adopts the innovative approach of evaluating machine learning model performance within the framework of business feasibility through a cost–benefit analysis that assesses the trade-off balance between performance metrics and the operational costs of hosting and deploying these models.

Although shown to provide noteworthy advances across various NLP tasks, the applicability of state-of-the-art generative language models for text classification tasks in zero-shot settings has been limited, especially for specialised domains. Large GLMs, such as GPT-3 (Brown et al. 2020) have been shown to perform robustly in few-shot learning tasks. However, they are less successful at zero-shot learning with significantly reduced performances on natural language inference tasks such as question answering and reading comprehension. In the absence of few-shot exemplars, it is possible that these models underperform on prompts that are dissimilar in format to that of their pre-training data (Wei et al. 2022).

Very few legal and contract datasets are publicly available. It is therefore unlikely that the OpenAI GPT family of models would have been trained on sufficient amounts of this type of niche data. This limits the chances of a multi-purpose large language model achieving zero-shot performance scores remotely competitive with those of smaller models trained exclusively on specialised legal datasets.

However, the fact that GPT-4 achieves an overall accuracy of 67.2% on the previously unseen CUAD-SL dataset, as well as outperforming the other benchmark models on particular clause categories, indicates the potential of these generative models to adapt to new tasks. This was achieved by a simple process of prompting through well-structured natural language descriptions. We believe this points the way to possible new hybrid zero-shot and meta-learning strategies that can leverage additional support strategies, such as weak supervision, clustering or topic modelling, to boost their performance on unseen categories. We briefly assess some of these potential approaches in the Conclusion and Future Work section.

The better performances of the other models highlights the benefit of task-targeted training when we are applying machine learning to domain-specific use cases such as legal clause classification. The shallow SGD classifier performs robustly when compared to the larger deep learning models. With an average accuracy performance of 83.7% across all classes, it outperforms OpenAI's Curie-FT model (81.7%) and the BERT transformer model (78.9%) and measures up comparably to the OpenAI Ada-FT's 83.8% performance on the CUAD-SL clause task. Bear in mind the vast differences in size and complexity of the model architectures.

There are challenges when it comes to adapting deep learning neural classifiers for legal domain tasks. Legal text differs in structure from the common, everyday language that is used in Wikipedia and other mass sources of text data that are used to train these state-of-the-art deep learning neural models. This complex and technical "legalese" can be a major barrier to Machine Learning fully understanding contracts, as no existing NLP-driven language models or off-the-shelf solutions can read legalese coherently (Lawgeex 2018). In the pursuit of adequately capturing legal domain-specific feature representations, proprietary Legal Language Processing (LLP) and Legal Language Understanding (LLU) models are likely to be the preferred strategies going forward. This approach has already yielded success in the text mining and information retrieval of unstructured biomedical data with the development of Biomedical Natural Language Processing (BioNLP) (Q. Chen et al. 2019; Zhang et al. 2019).

We are now entering an era where AI is likely to make its biggest impact in the business and industrial world. For businesses, machine learning is a particularly expensive investment. It requires many highly skilled professionals, a sophisticated infrastructure, and a governance and mindset that very few companies actually have.

Predictive performance metrics are the currency of academic research papers. However, when considering a business context, we have to adopt a business-centric analysis that transcends the evaluation of raw predictive performance and accounts for the day-to-day cost implications of these model choices. Predictive performance is an important and reliable measure, but it might not compensate for increased costs or measures of business impact. Cost–benefit and cost-effectiveness analyses are simple but powerful tools to allow business decision-makers to systematically compare the pros and cons of alternative models in terms of the jobs that they are going to be used for. Ultimately the best model will always be the one that serves its purpose best and impacts the business in a manner worthy of its costs.

Our benchmark evaluation shows that DeBERTa was the top performing model with an accuracy of 87.8% on the CUAD SL dataset. Framing this performance within a cost–benefit analysis reveals that under a low usage business scenario, DeBERTa is the best option with a running cost of £38 per day. However, depending on budgetary constraints, this picture changes when we analyse costs under a high usage scenario. In this situation, the cost of training, hosting and deploying DeBERTa could be as much as £134 per day which may be prohibitive depending on the size and budget constraints of certain businesses, especially since this is just one model for one specific task. If the end task for which the model is being deployed is lower risk and can accommodate a degree of flexibility within the performance accuracy, then an option such as Ada-FT operating at 83.8% and a cost of £20 under high usage conditions may be a more cost efficient alternative.

Having a complete picture of the balance between costs and model performance is essential for the role of structuring the decision-making process and strategic planning of the deployment of machine learning solutions at scale within a business. Finding the best trade off between costs and predictive power ultimately boils down to a simple decision in this particular legal use case scenario—is the business happy to pay roughly seven times more for a model that is 4% more accurate for the task of clause classification?

6 Conclusion & future work

As there will always be scenarios where it is too costly or impractical to obtain sufficient amounts of labelled data, strategies that facilitate label-efficient large scale text classification, such as zero-shot learning, weak supervision and data augmentation will be greatly in demand.

An obvious path of improvement would be to restructure the task from zero-shot to a few-shot text classification problem (Zhou et al. 2024). A proposed human-in-the loop method would involve a legal domain expert selecting representative class samples with their associated label for each clause and providing them in the prompt input. This is certainly achievable for smaller multi-class classification tasks but may challenge the prompt context token limit when we consider larger multi-class problems such as this study where the classifier has to differentiate between 23 different clause categories, some of which look very similar to a layperson (or large language model) but entail distinct legal obligations.

The generative and non-deterministic behaviour of GLMs may effectively rule them out as reliable and robust tools for text classification as there will always be the effects of randomness and hallucination to contend with. Where GLMs may be best leveraged for text classification is in the synthesis of new and additional training data through the augmentation or generation of existing base data.

The benefits of deploying GLMs as tools for data synthesis within the legal domain are manifold. Creating sufficient legal training data is associated with high labelling costs in terms of resources, time and finance. GLM-driven text generation can help overcome the training bottleneck by minimising the amount of data to be labelled and can be particularly useful in time critical scenarios where data has to be collected quickly.

Appendix 1

CUAD Restructuring into single label dataset

A sentence or paragraph can have multiple labels. Many of these labels are overlapping in terms of the specific conditions they are trying to control for. As such, there is an opportunity to simplify the CUAD dataset to reformat it as a one label classification problem. In doing so, we simplify the dataset but also make it more robust and a fairer benchmark for future classification research studies.

For the sake of full transparency, we map the original multi-label CUAD dataset labels to the new single label set-up, providing evidence and clause definitions to justify each decision:

Transfer restrictions

Original CUAD Label	New Label
Anti-assignment	Transfer Restrictions
Non-Transferable License	

Justification: Anti-Assignment is defined as a provision which prevents one of both of the parties to a contract from assigning or transferring some or all of their respective obligations or rights to a third party. The purpose of an Anti- Assignment clause is to protect the expectations and interests of the original parties to a contract by preventing unknown or undesirable assignees or obligators from affecting the performance or enforcement of the contract (Boessel n.d.) Lawinsider defines Non-Transferable Licence as Licences which cannot be assigned or sublicensed because of their terms or pursuant to law. It aims to protect the licensor's intellectual property, quality control, or contractual interests by preventing the licensee from using the licence in ways that the licensor did not authorise (Non-Transferable Licenses – Definition n.d.). As both clauses deal with the prevention of the transfer of certain contractual rights, a decision was made to merge them under the catch-all umbrella term of 'Transfer Restrictions'.

Liability limit

Original CUAD Label	New Label
Cap On Liability	Liability Limit
Uncapped Liability	

Justification: According to Ashurst, a Cap On Liability clause outlines the maximum amount payable in damages by the parties to one another upon an event of default, restricting the losses that can be recovered or the remedies available, or imposing time limits on when claims can be made (Sneddon n.d.). This can be contrasted with an Uncapped Liability clause, which neither limits the amount nor types of damages one party can claim from the other in a case of dispute. As such, both types of clause have been grouped together under the label 'Liability Limit' as they address the extent of the contracting parties' liability to one another in an event of a breach, negligence, or other wrongdoing.

Contract details

Original CUAD Label	New Label
Document Name	Contract Details
Parties	
Agreement Date	
Effective Date	
Expiration Date	
Renewal Term	

Justification: Upcounsel describes a Document Name clause as a descriptive or specific heading that outlines the purpose of the contract, and a Parties clause as a section that stipulates the legal entitles or individuals that agree to be bound by its terms (Parts of a Contract: Everything You Need to Know 2020). Agreement Date, Effective Date and Expiration Date outline the date on which an agreement is signed and the parties are bound by the terms of the contract, the date on which a contracts obligations are commenced, and the date upon which a contract expires according to its terms, respectively (Commercial and Technological Contracts Mythbuster—Backdating a Contract 2020). A Renewal Term clause is expressed by Practical Law as being a clause which stipulates the conditions and procedures for extending the duration of a contract beyond its initial term (General Contract Clauses: Term and Termination n.d.). These clauses can be considered more as a collection of data points and have been combined under the umbrella label of 'Contract Details' as together they are regarded as the key elements of a contract which must be present for an agreement to be legally enforceable.

Insurance and liquidated damages

Original CUAD Label	New Label
Insurance	Insurance and Liquidated Damages
Liquidated Damages	

Justification: An Insurance clause stipulates the limitations of liability policy conditions and general liability risks an insurance provider covers for the duration of the contract. A Liquidated Damages clause can be described as a remedies clause which requires a party in breach of a contract to pay a pre-determined fixed-amount, or an amount based on a pre-determined formula, as compensation to the non-breaching party for failure to meet their contractual obligations (Using Contractual Risk Allocation Provisions to Minimize Risk and Maximize Reward n.d.). These clauses contend with liability and loss among the parties to the contract and are thus jointly represented under a new label entitled 'Insurance and Liquidated Damages' for this reason. The two original labels have been integrated into one overarching term that references both label names so as to appreciate the intricacies of each clause.

IP Ownership

Original CUAD Label	New Label
IP Ownership Assignment	IP Ownership
Joint IP Ownership	
Source Code Escrow	

Justification: IP Ownership Assignment is a provision that stipulates how the ownership and rights to IP is to be transferred from the inventor or owner to another entity (Miller 2022). Joint IP Ownership is a clause which regulates the scope of IP which is jointly-owned by two or more parties in terms of allocation of the ownership shares or percentages among the co-owners, their rights and responsibilities, the procedures and mechanisms for resolving disputes regarding the joint IP, and the duration and termination of the joint IP. A Source Code Escrow clause outlines the arrangement between the licensor and a licensee of software for the former to deposit a copy of the software's source code with a third-party escrow agent to hold in trust for the latter (Source Code Escrow n.d.). Consequently, these clauses have been grouped together under the label 'Ip Ownership' because they address how IP rights created or used throughout the duration of a contract are to be allocated between parties.

Termination rights

Original CUAD Label	New Label
Notice Period To Terminate Renewal	Termination Rights
Termination For Convenience	

Justification: A Notice Period to Terminate Renewal clause specifies the timeframe in which one or both of the parties to the contract must inform the other of their desire to end the contract at the end of its current term, rather than renewing it for another period. A Termination for Convenience clause is a contractual right to end an agreement without cause or liability, subject to certain conditions and limitations (Rishi 2020b). As these clauses stipulate the terms upon which a contract can be brought to an end, it was agreed to categorise them under the label 'Termination Rights'.

Exclusivity

Original CUAD Label	New Label
Non-Disparagement	Exclusivity
Non-compete	
Exclusivity	
No-Solicit Of Customers	
No-Solicit Of Employees	

Justification: According to Practical Law, a Non-Disparagement clause prohibits one or both parties from making negative or derogatory statements about the other party, their products, services, reputation, or interests, both during and following the contract term. As such, it is referred to as a Restrictive Covenant due to its constricting impact on the actions of contracting parties (Non-Disparagement Provision n.d.). Non-compete and non-solicit provisions are also examples of restrictive covenants; a non-compete clause prevents one party from engaging in certain activities or businesses that compete with or harm the other party, typically for a specified duration and within a defined geographic area (The Noncompete Clause Explained 2021), and No-Solicit of Customers and No-Solicit of Employees clauses prohibit one party from directly or indirectly soliciting the customers or employees of another party during a certain time or specified territory, respectively (Non-Solicitation n.d.). The aforementioned clauses can be captured under the more generic Exclusivity clause, which is a provision which restricts one or both parties to the agreement from soliciting offers or negotiating with a third parties during or after the term of the contract, and thus is the reason why they have been merged under this label accordingly.

Licence terms

Original CUAD Label	New Label
Irrevocable Or Perpetual License	Licence Terms
Affiliate Licence-Licensee	
Affiliate Licence-Licensor	

Justification: An irrevocable or perpetual license clause grants a non-terminable and non-revocable right to exploit IP or other asset, subject to the contract's conditions (What are Perpetual, Irrevocable, Royalty-free Licences n.d.). An Affiliate License-Licensee gives the licensee the right to use, distribute or sub-license certain IP or other assets that are owned or controlled by the licensor's affiliates, whereas an Affiliate License-Licensor clause grants or restricts the rights of the parties to sublicense, assign, or otherwise transfer their license to use, distribute, or modify a certain product, service, or intellectual property to their affiliates (Licensor vs Licensee—What's the Difference n.d.). All three of the aforementioned clauses have been brought together under the catch-all label entitled 'License Terms' due to the fact that they deal with the rights and restrictions regarding the use of IP and other assets.

Licence scope

Original CUAD Label	New Label
Licence Grant	Licence Scope
Unlimited/All-You-Can-Eat-License	

Justification: A Licence Grant is a provision which specifies the terms and conditions under which the licensor grants the licensee the right to use the IP in a way that would otherwise be an infringement had the licence not been granted (Guidance on Licensing Intellectual Property n.d.). An Unlimited/All-You-Can-Eat-License grants a right to use IP without limitation in the same manner as the true owner (What does Limited or Unlimited Mean in a License 2017). These clauses have been combined under the label 'Licence Scope' because they outline the extent of the rights afforded by a license.

Glossary of legal clauses

- **Affiliate License-Licensee –** A person to whom an affiliate licence is granted.

  A contractual provision which grants the licensee the right to use, distribute, or sublicense certain IP or other assets of the licensor that are owned or controlled by the latter's affiliates.

- **Affiliate License-Licensor –** A party that grants an affiliate licence to another (General Contract Clauses: Audit Rights n.d.).

  A contractual provision that grants or restricts the rights of the parties to sub-license, assign, or otherwise transfer their license to use, distribute, or modify a certain product, service, or intellectual property to their affiliates.

- **Agreement Date –** The date on which an agreement is signed and the parties are bound by the terms of the contract (Commercial and Technological Contracts Mythbuster—Backdating a Contract 2020).

  This may also indicate the Effective Date of the contract if different from the Agreement Date.

- **Anti-Assignment –** A provision which prevents one of both of the parties to a contract from assigning or transferring some or all of their respective obligations or rights to a third party (Boessel n.d.).

  This protects the expectations and interests of the original parties to a contract by preventing unknown or undesirable assignees or obligators from affecting the performance or enforcement of the contract.

- **Audit Rights –** A standard clause that stipulates the audit rights and obligations of the contracting parties (General Contract Clauses: Audit Rights n.d.).

  More specifically, it is a provision which grants one party the right to inspect, examine or review the records, books, accounts, or other information of another party, typically for the purpose of verifying compliance, accuracy, performance, or quality. An Audit Rights clause may specify the details of the audit, as well as the consequences of non-compliance or discrepancies.

- **Cap On Liability –** A standard clause limiting the liability of contracting parties to one another in an event of default (Sneddon n.d.).

  The purpose of a cap on liability clause is to allocate the risk of potential losses between the parties and to provide certainty and predictability in case of a dispute.

- **Change Of Control –** A provision in an agreement which grants a party a specific right or entitlement in the event of a change in ownership or management of the other party to the agreement (Change of Control Clause n.d.).

- **Competitive** Restriction **Exception** – An exception to a non-compete provision.

  A contractual provision or agreement that allows one or more parties to engage in certain activities or transactions that would otherwise be prohibited or limited by a non-compete, non-solicitation, non-disclosure, or other restrictive covenant. The purpose of the clause is to stipulate circumstances where the parties agree that a degree of competition or disclosure outweigh the potential harms or risks of breaching the covenant.

- **Covenant Not To Sue** – A contractual provision that restricts a party from claiming damages from the other, usually in exchange for some form of compensation or benefit (Kagan 2022).

- **Document Name** – Establishes the purpose of the contract (Parts of a Contract 2020).

  It outlines the title or name of the contract, and is usually found at the beginning of the document, either as part of the introductory paragraph or as a separate heading. The Document Name should be descriptive and specific enough to capture the nature and the scope of the agreement.

- **Effective Date**—The date on which a contracts obligations are commenced (Commercial and Technological Contracts Mythbuster—Backdating a Contract 2020).

  This may be the same as the date of signing, a future date triggered by a certain event, or a different date agreed by the parties.

- **Exclusivity** – An agreement which prevents one party from soliciting offers or negotiating with a third party for a specific period of time or in a specified territory (Exclusivity Agreement n.d.).

  The purpose of this clause is to protect the interests and investments of the parties, to ensure loyalty and commitment, and prevent the dilution or diversion of market share, customers, or intellectual property.

- **Expiration Date** – The date upon which a contract expires according to its terms.

  It may also include conditions or consequences for terminating, renewing, or extending the contract or agreement before or after the expiration date.

- **Governing Law—**A clause stipulating the legal jurisdiction the parties have nominated to govern the performance and interpretation of their agreement and whose courts will determine any potential disputes arising under it (Governing Law n.d.).

- **Insurance** – A provision which stipulates the limitations of liability policy conditions and general liability risks an insurance provider covers (Insurance Clause n.d.).

- **IP Ownership Assignment** – An agreement to transfer ownership and all rights of the intellectual property from the creator to another entity (Miller 2022).

  The purpose of the clause is to clarify and confirm who owns the IP, who can use it, and under what terms and conditions.

- **Irrevocable Or Perpetual License** – Irrevocable licences stipulate whether the licence cannot be revoked or terminated in perpetuity or if the licence cannot be revoked, other than subject to the term and termination provisions in the agreement/Perpetual licences stipulate whether the licence is 'never-ending', of an 'indefinite duration' until terminated in accordance with its terms or for a long fixed-term period (What are Perpetual, Irrevocable, Royalty-free Licences n.d.).

  It is a provision that grants a non-terminable, non-expiring, and non-revocable right to use, copy, distribute, modify, or otherwise exploit IP or another asset, subject to the terms and conditions of the contract.

- **Joint Ip Ownership** – A clause which regulates the use of intellectual property which is jointly-owned by two or more parties (Gledhill 2022).

  The provision may address the scope of the IP that is subject to joint ownership, allocation of the ownership shares or percentages among the co-owners, the rights and responsibilities of the co-owners of the joint IP, the procedures and mechanisms for resolving disputes relating to the joint IP, and the duration and termination of the joint IP.

- **License Grant** – An agreement between the IP right owner and a third party that permits the latter to use the IP in a way that would otherwise be an infringement had the licence not been granted (Guidance on Licensing Intellectual Property n.d.).

  It is a provision that specifies the terms and conditions under which the licensor grants the licensee the right to use, access, or exploit a certain IP.

- **Liquidated Damages** – A type of exclusive remedies clause requiring a party in breach of a contract to pay a pre-determined fixed-amount, or an amount based on a pre-determined formula, as compensation to the non-breaching party for failure to meet their contractual obligations (Using Contractual Risk Allocation Provisions to Minimize Risk and Maximize Reward n.d.)

- **Minimum Commitment** – A minimum obligation or requirement clause stipulating the base terms in relation to various aspects of an agreement.

  It obligates one or both parties to perform a certain level of activity, output, or expenditure over a specified period of time, regardless of changes in demand, market conditions, or other factors in order to either secure a stable and predictable revenue stream, ensure a sufficient return on investment, or incentivise performance and loyalty.

- **Most Favored Nation –** A contractual clause that requires a country to provide the same trade terms to all trading partners (Kenton 2022).

  In other words, it is a provision that grants one party the same or better terms and conditions as those given to any other party in a similar or comparable situation.

- **No-Solicit Of Customers –** A restrictive covenant that prohibits one party from directly or indirectly soliciting the customers of another party during a specified period of time and within a defined geographic area (Non-Solicitation n.d.).

  Its purpose is to protect the goodwill, reputation, and competitive advantage of the party who has established a relationship with the customers or clients, and to prevent the other party from unfairly exploiting or interfering with that relationship.

- **No-Solicit Of Employees** – A covenant that prohibits one party from directly or indirectly soliciting the employees of another party during a specified period of time and within a defined geographic area (Non-Solicitation n.d.).

  Its purpose is to protect the employer's investment in its human capital, prevent the loss of valuable skills and knowledge, and avoid disruption to its business operations.

- **Non-Compete** – A legally enforceable term in an employment contract which prevents an employee from working for the competitors of their previous employer for a specific period of time and within a defined geographic area following resignation or termination of employment (The Noncompete Clause Explained 2021).

  A non-compete clause in a contract is a provision that restricts one party from engaging in certain activities or businesses that compete with or harm the interests of another party, usually for a specified period of time and within a defined geographic area. The purpose of a non-compete clause is to prevent the other party from using the skills, knowledge, or contacts gained from the relationship to unfairly compete or harm the business of the first party.

- **Non-Disparagement –** A clause which restricts what an employee or an employer can say about one another during and following the period of employment (Non-Disparagement Provision n.d.).

  In other words, it is a provision in a contract that prohibits one or both parties from making negative or derogatory statements about the other party, their products, services, reputation, or interests, either during or after the contractual relationship. The purpose of a non-disparagement clause is to protect the goodwill, reputation, and business interests of the parties from harm caused by unfair or malicious criticism, defamation, or slander. Non-disparagement clauses are not limited to employment agreements, but may be included in other types of contracts, such as settlement agreements, severance agreements, confidentiality agreements, partnership agreements, vendor agreements, or customer agreements.

- **Non-Transferable License –** Licences which cannot be assigned or sublicensed because of their terms or pursuant to law (Non-Transferable Licenses – Definition n.d.).

  It aims to protect the licensor's intellectual property, quality control, or contractual interests by preventing the licensee from using the licence in ways that the licensor did not authorise.

- **Notice Period To Terminate Renewal –** A clause stipulating the required notice period to terminate the renewal of a contract.

  A notice period to terminate renewal clause in a contract is a provision that specifies how much time in advance one or both parties must give to the other if they wish to end the contract at the end of its current term, rather than renewing it for another period. A notice period to terminate renewal clause is often used in contracts that have automatic or periodic renewal options, such as leases, subscriptions, service agreements, or licenses. The purpose of a notice period to terminate renewal clause is to provide certainty and clarity for both parties about their rights and obligations regarding the continuation or termination of the contract, and to avoid disputes or misunderstandings that may arise from implied or oral agreements.

- **Parties –** The names of the parties involved in a contract (Parts of a Contract 2020).

  A Parties clause in a contract is a section that identifies and defines the legal entities or individuals who are entering into the agreement and are bound by its terms and conditions.

- **Post-Termination Services –** Termination clauses stipulate which contractual rights are to continue or end when the contract is terminated (Consequences of Termination Clause—Post-Termination Rights n.d.).

  A post-termination services clause specifies the obligations and rights of the parties after the termination of the contract, especially regarding the continuation or cessation of any services that were provided or received under the contract.

- **Price Restrictions –** A clause stipulating product and service pricing limits (Price Restrictions n.d.).

  A price restrictions clause in a contract is a provision that limits or regulates the amount, method, or timing of payment for goods or services exchanged between the parties.

- **Renewal Term –** A clause stipulating the renewed length of a contract upon expiration of its initial term (General Contract Clauses: Term and Termination n.d.).

  It is a provision that specifies the conditions and procedures for extending the duration of the contract beyond its original expiration date.

- **Revenue/Profit Sharing –** An agreement between two parties where one party must pay a percentage of the profits or revenues received to the other for their contribution to the business (Revenue Sharing Agreement n.d.).

  It specifies how the parties to an agreement will divide the income or earnings generated by a joint venture, project, partnership, or other collaborative arrangement. The clause may define the sources, methods, and timing of revenue or profit calculation and distribution, as well as the rights and obligations of each party regarding accounting, reporting, auditing, taxes, and disputes.

- **Rofr/Rofo/Rofn – Right of First Refusal (ROFR)** is a contractual right given to non-selling shareholders to either accept or refuse an offer from a selling shareholder after the selling shareholder has received a third party offer for its shares. **Right of First Offer (ROFO)** gives non-selling shareholders the right to make an offer for the selling shareholder's shares before the selling shareholder can offer its shares to third-parties (Shareholder's Agreements 2017). A **Right of First Negotiation (ROFN)** requires the grantor to negotiate with the holder for a transaction during a certain period of time (Right of First Negotiation, Offer, and Refusal n.d.).

  Rofr gives one party the option to match or exceed any offer that another party receives or makes for a certain asset, transaction, or opportunity, before the other party can accept or pursue it. Rofo gives one party (the holder) the opportunity to make an offer to buy, sell, lease, or otherwise acquire or dispose of a specified asset, property, or interest before the other party (the grantor) can solicit or accept offers from third parties. Rofn grants one party the opportunity to negotiate the terms of a potential deal with another party before the other party can solicit or accept offers from third parties.

- **Source Code Escrow –** An agreement between the licensor and a licensee of software for the licensor to deposit a copy of the software's source code with a third-party escrow agent (Source Code Escrow n.d.).

  It requires a software developer or vendor to deposit a copy of the source code of a software product or application with a third-party escrow agent, who holds it in trust for a software licensee or customer. The purpose of a source code escrow clause is to protect the licensee or customer's interests in circumstances where the developer or vendor goes bankrupt, breaches the license agreement, fails to provide adequate support or maintenance, or otherwise becomes unable or unwilling to fulfil its obligations regarding the software. In such scenarios, the escrow agent can release the source code to the licensee or customer to be modified or maintained.

- **Termination For Convenience –** A contractual right to end an agreement without cause by providing notice of termination to the other party (Rishi 2020a, b).

  It allows one or both parties to end the agreement without cause or liability, subject to certain conditions and limitations.

- **Third Party Beneficiary –** A clause which allows contracting parties to specify exceptions to which third-party beneficiaries can benefit from and enforce the contract (General Contract Clauses: Third-Party Beneficiaries n.d.).

  The provision grants rights or benefits to a person or entity that is not a party to the contract, but is intended by the contracting parties to receive some advantage from the contract's performance. The clause usually specifies the identity or class of the third party beneficiary, the nature and extent of the rights or benefits conferred, and the conditions or limitations for enforcing those rights or benefits.

- **Uncapped Liability –** Liability without limit.

  An uncapped liability clause stipulates that there is no limit on the amount or type of damages that one party can claim from another in the event of a breach, negligence, or other wrongdoing. It exposes the liable party to the risk of paying the full extent of the actual or potential losses suffered by the other party, regardless of whether they are direct, indirect, consequential, punitive, or otherwise.

- **Unlimited/All-You-Can-Eat-License –** A right to use intellectual property without limitation in the same manner as a true owner of the intellectual property (What does Limited or Unlimited Mean in a License 2017).

  An unlimited licence clause grants one party the right to use, copy, distribute, modify, or otherwise exploit the intellectual property, software, data, or other assets of another party without any restrictions on the scope, duration, purpose, or territory of the licence. It may also waive any fees, royalties, or compensation for the licensor, and may exclude any warranties, liabilities, or indemnification obligations for the licensee.

- **Volume Restriction –** A clause stipulating that the quantity of shares or stock shall not exceed a certain amount for a specified period of time.

  A volume restriction clause limits the quantity or value of goods or services that one party can buy, sell, supply, or receive from another party, usually within a specified period or market.

- **Warranty Duration –** A clause stipulating the length of a warranty period.

  A warranty duration clause specifies how long a party's warranty obligations last. A warranty is a promise or guarantee that a product or service meets certain standards of quality, performance, or functionality, or that a party has certain rights or authority to enter into a contract. A warranty duration clause may also define the conditions, limitations, and remedies for any breach of warranty.

Appendix 2

Zero shot prompt structure for OpenAI – GPT-4 8 K model

The prompt instructs the model to classify the input text using the list of 23 possible labels from the CUAD-SL dataset. The prompt structure is as follows:

Context:

[Clause text from CUAD-SL inputted here]
Prompt: You are a lawyer. You will classify the Context using one of the following labels. Only respond with the label and nothing else.
Labels:
Governing Law
Audit Rights
Liability Limit
….
….
Competitive Restriction Exception

Appendix 3

Table 4 Recall and precision results across all labels of the CUAD-SL dataset for the OpenAI, TFI + SGD and BERT models

Recall	GPT4	AdaFT	CurieFT	SGD	DeBERTa
Contract Details	61%	88%	86%	94%	95%
Transfer Restrictions	70%	87%	87%	84%	90%
Liability Limit	86%	90%	81%	94%	92%
Exclusivity	46%	82%	87%	75%	86%
Insurance and Liquidated Damages	84%	90%	83%	91%	94%
Audit Rights	94%	91%	91%	96%	94%
Revenue/Profit Sharing	82%	87%	84%	87%	93%
Ip Ownership	70%	81%	83%	87%	85%
Governing Law	100%	99%	96%	99%	99%
Minimum Commitment	58%	80%	85%	75%	87%
Post-Termination Services	41%	78%	72%	82%	82%
License Scope	54%	66%	62%	60%	64%
Rofr/Rofo/Rofn	27%	78%	75%	84%	81%
Termination Rights	94%	83%	80%	87%	83%
Change Of Control	63%	72%	74%	67%	79%
License Terms	32%	58%	59%	56%	63%
Warranty Duration	80%	91%	78%	85%	93%
Volume Restriction	34%	68%	72%	53%	72%
Covenant Not To Sue	14%	86%	82%	68%	86%
Competitive Restriction Exception	57%	23%	20%	13%	50%
Third Party Beneficiary	92%	83%	83%	33%	100%
Most Favored Nation	73%	82%	91%	9%	82%
Price Restrictions	86%	43%	57%	0%	57%
Precision	GPT4	AdaFT	CurieFT	SGD	DeBERTa
Contract Details	83%	95%	96%	89%	94%
Transfer Restrictions	79%	87%	89%	84%	87%
Liability Limit	93%	92%	96%	93%	93%
Exclusivity	80%	69%	63%	70%	79%
Insurance and Liquidated Damages	96%	95%	97%	94%	95%
Audit Rights	92%	96%	97%	91%	97%
Revenue/Profit Sharing	91%	90%	93%	85%	90%
Ip Ownership	71%	89%	85%	80%	88%
Governing Law	97%	100%	97%	98%	99%
Minimum Commitment	84%	84%	71%	72%	82%
Post-Termination Services	87%	76%	75%	75%	79%
License Scope	35%	73%	69%	65%	70%
Rofr/Rofo/Rofn	97%	79%	78%	84%	88%
Termination Rights	33%	79%	72%	80%	86%
Change Of Control	76%	81%	76%	72%	70%
License Terms	14%	77%	76%	67%	79%
Warranty Duration	79%	85%	86%	89%	85%
Volume Restriction	62%	78%	83%	74%	88%
Covenant Not To Sue	100%	91%	100%	92%	86%

Table 4 (continued)

Competitive Restriction Exception	20%	54%	75%	40%	52%
Third Party Beneficiary	85%	100%	91%	100%	92%
Most Favored Nation	89%	82%	71%	100%	75%
Price Restrictions	19%	75%	57%	-	80%

Table only showing results for DeBERTa as this was the top performer from the BERT-based trials

References

Bayer M, Kaufhold MA, Reuter C (2022) A survey on data augmentation for text classification. ACM Comput Surv. https://doi.org/10.1145/3544558

Branavan SRK, Chen H, Zettlemoyer LS, Barzilay R (2009) Reinforcement learning for mapping instructions to actions. ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP. Proceedings of the Conf, pp 82–90. https://doi.org/10.3115/1687878.1687892

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Amodei D (2020) Language models are few-shot learners. In: Advances in neural information processing systems 33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: The muppets straight out of law school. Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020, pp 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz DM, Aletras N (2022) LexGLUE: a benchmark dataset for legal language understanding in english. Proc Ann Meet Assoc Comput Linguist 1:4310–4330. https://doi.org/10.2139/ssrn.3936759

Chen DL, Mooney RJ (2011) Learning to interpret natural language navigation instructions from observations. Proc AAAI Conf Artif Intell 25(1):859–865. https://doi.org/10.1609/aaai.v25i1.7974

Chen H, Wu L, Chen J, Lu W, Ding J (2022) A comparative study of automated legal text classification using random forests and deep learning. Inf Process Manage 59(2):102798. https://doi.org/10.1016/j.ipm.2021.102798

Chen Q, Peng Y, Lu Z (2019) BioSentVec: creating sentence embeddings for biomedical texts. 2019 IEEE International Conference on Healthcare Informatics, ICHI. https://doi.org/10.1109/ICHI.2019.8904728

Chiticariu L, Li Y, Reiss FR (2013) Rule-based information extraction is dead! Long live rule-based information extraction systems!. EMNLP 2013 - 2013 conference on empirical methods in natural language processing, Proceedings of the Conference, pp 827–832. https://aclanthology.org/D13-1079. Accessed 16 Jul 2024

Clavié B, Alphonsus M (2021) The unreasonable effectiveness of the baseline: discussing SVMs in legal text classification. In: Frontiers in artificial intelligence and applications, vol 346, pp 58–61. https://doi.org/10.3233/FAIA210317

Collier K, Brand M, Pramod N (2019) Models of enterprise intelligence. Thoughtworks. https://www.thoughtworks.com/en-gb/insights/articles/intelligent-enterprise-series-models-enterprise-intelligence. Accessed 18 Jul 2024

Costa YDR, Oliveira H, Nogueira V, Massa L, Yang X, Barbosa A, Oliveira K, Vieira T (2023) Automating petition classification in Brazil's legal system: a two-step deep learning approach. Artif Intell Law. https://doi.org/10.1007/s10506-023-09385-4

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American chapter of the association for computational linguistics: human language technologies - proceedings of the conference, vol 1, pp 4171–4186

Diab S (2019) Optimizing stochastic gradient descent in text classification based on fine-tuning hyper-parameters approach. A case study on automatic classification of global terrorist attacks. ArXiv. https://arxiv.org/abs/1902.06542. Accessed 18 Jul 2024

Ding T, Chen T, Zhu H, Jiang J, Zhong Y, Zhou J, Wang G, Zhu Z, Zharkov I, Liang L (2024) The efficiency spectrum of large language models: an algorithmic survey. ArXiv. https://doi.org/10.48550/arXiv.2312.00678. Accessed 15 Jul 2024.

Dumais S, Piatt J, Heckerman D, Sahami M (1998) Inductive learning algorithms and representations for text categorization. Int Conf Inf Knowl Manag Proc 148–155. https://doi.org/10.1145/288627.288651

Gera A, Halfon A, Shnarch E, Perlitz Y, Ein-Dor L, Slonim N (2022) Zero-shot text classification with Self-training. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp 1107–1119. https://doi.org/10.18653/v1/2022.emnlp-main.73

Glavas G, Nanni F, Ponzetto SP (2016) Unsupervised text segmentation using semantic relatedness graphs. *SEM 2016 - 5th Joint Conference on Lexical and Computational Semantics, Proceedings, pp 125–130. https://doi.org/10.18653/v1/s16-2016

Graham SG, Soltani H, Isiaq O (2023) Natural language processing for legal document review: categorising deontic modalities in contracts. Artif Intell Law. https://doi.org/10.1007/s10506-023-09379-2

Gu J, Wang Y, Chen Y, Cho K, Li VOK (2018) Meta-Learning for Low-Resource Neural Machine Translation. ArXiv, abs/1808.08437. https://doi.org/10.48550/arXiv.1808.08437

Hausladen CI, Schubert MH, Ash E (2020) Text classification of ideological direction in judicial opinions. Int Rev Law Econ 62:105903. https://doi.org/10.1016/j.irle.2020.105903

Hendrycks D, Burns C, Chen A, Ball S (2021) CUAD: an expert-annotated NLP dataset for legal contract review. ArXiv. https://doi.org/10.48550/arXiv.2103.06268

Huang PS, Wang C, Singh R, Yih WT, He X (2018) Natural language to structured query generation via meta-learning. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 2:732–738. https://doi.org/10.18653/v1/n18-2115

Huggingface (n.d.) from https://huggingface.co/ Accessed 16 Jul 2024

Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: Nédellec C and Rouveirol C (eds.). Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics. Springer Berlin Heidelberg, vol 1398, pp 137–142. https://doi.org/10.1007/s13928716

Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference 2:427–431. https://doi.org/10.18653/v1/e17-2068

Laskar MTR, Bari MS, Rahman M, Bhuiyan MAH, Joty S, Huang JX (2023) A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp 431–469. https://doi.org/10.18653/v1/2023.findings-acl.29

Lawgeex Blog (2018) Comparing the performance of artificial intelligence to human lawyers in the review of standard business contracts. https://www.lawgeex.com. Accessed 16 Jul 2024

Li Q, Peng H, Li J, Xia C, Yang R, Sun L, Yu PS, He L (2020) A survey on text classification: from shallow to deep learning. ArXiv, abs/2008.0. http://arxiv.org/abs/2008.00364. Accessed Jul 18 2024

Lu Q, Dou D, Nguyen TH (2022) ClinicalT5: a generative language model for clinical text. Findings of the association for computational linguistics: EMNLP 2022, pp 5465–5472. https://doi.org/10.18653/v1/2022.findings-emnlp.398

McCallum A, Nigam K (1998) A comparison of event models for naive bayes text classification. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp 41–48. 10.1.1.46.1529. https://cdn.aaai.org/Workshops/1998/WS-98-05/WS98-05-007.pdf. Accessed 17 Jul 2024

Meng Y, Zhang Y, Huang J, Xiong C, Ji H, Zhang C, Han J (2020) Text classification using label names only: a language model self-training approach. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp 9006–9017. https://doi.org/10.18653/v1/2020.emnlp-main.724

Microsoft (2023) Azure openAI aervice. Microsoft. https://azure.microsoft.com/en-us/products/ai-services/openai-service/. Accessed 15 Jul 2024

Moradi M, Blagec K, Haberl F, Samwald M (2021) GPT-3 Models are poor few-shot learners in the biomedical domain. http://arxiv.org/abs/2109.02555. Accessed 17 Jul 2024

Nallapati R, Manning CD (2008) Legal docket-entry classification: where Machine Learning stumbles. EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL, pp 438–446. https://aclanthology.org/D08-1046 Accessed 14 Jul 2024

OpenAI (n.d.) from https://openai.com/ Accessed 22 Jul 2024

Price E, Masood A, Aroraa G (2021) Azure machine learning. Hands-on Azure Cogn Serv. https://doi.org/10.1007/978-1-4842-7249-7_10

Prusa J, Khoshgoftaar TM, Seliya N (2016) The effect of dataset size on training tweet sentiment classifiers. Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015, pp 96–102. https://doi.org/10.1109/ICMLA.2015.22

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners | enhanced reader. OpenAI Blog 1(8):9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed 15 Jul 2024

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1–67. https://doi.org/10.48550/arXiv.1910.10683. Accessed 20 Jul 2024

Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: Unanswerable questions for SQuAD. ACL 2018 - 56th annual meeting of the association for computational linguistics, proceedings of the conference (Long Papers), 2:784–789. https://doi.org/10.18653/v1/p18-2124

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuad: 100,000+ questions for machine comprehension of text. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, pp 2383–2392. https://doi.org/10.18653/v1/d16-1264

Rishi A (2020a) The complete list of standard clauses to check before signing a contract. Spotdraft. https://www.spotdraft.com/blog/standard-clauses-to-check-contract. Accessed 19 Jul 2024

Schopf T, Braun D, Matthes F (2022) Evaluating unsupervised text classification: zero-shot and similarity-based approaches. ACM International Conference Proceeding Series, pp 6–15. https://doi.org/10.1145/3582768.3582795

Tavor AA, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, Tepper N, Zwerdling N (2020) Do not have enough data? Deep learning to the rescue! AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, pp 7383–7390. https://doi.org/10.1609/aaai.v34i05.6233

The Atticus Project (n.d.) from https://www.atticusprojectai.org/. Accessed 5 Jul 2024

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019) SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In: Advances in Neural Information Processing Systems. Curran Associates Inc., vol 32

Wang S, Manning CD (2012) Baselines and bigrams: Simple, good sentiment and topic classification. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, vol 2. Association for Computational Linguistics, Korea. pp 90–94. https://aclanthology.org/P12-2018. Accessed 18 Jul 2024

Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV (2022) Finetuned language models are zero-shot learners. In: ICLR 2022 - 10th International Conference on Learning Representations https://doi.org/10.48550/arXiv.2109.01652

Wilcox A, Hripcsak G (1999) Classification algorithms applied to narrative reports. Proceedings / AMIA ... Annual Symposium. AMIA Symposium, pp 455–459

Yan G, Li Y, Zhang S, Chen Z (2019) Data augmentation for deep learning of judgment documents. In: Cui Z, Pan J, Zhang S, Xiao L, Yang J (Eds.). Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): LNCS. Springer International Publishing, vol 11936 pp 232–242. https://doi.org/10.1007/978-3-030-36204-1_19

Yang Y, Liu X (1999) A re-examination of text categorization methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR, pp 42–49. https://doi.org/10.1145/312624.312647

Zhang Y, Chen Q, Yang Z, Lin H, Lu Z (2019) Biowordvec, improving biomedical word embeddings with subword information and MeSH. Sci Data 6(1):1–9. https://doi.org/10.1038/s41597-019-0055-0

Zheng L, Guha N, Anderson BR, Henderson P, Ho DE (2021) When does pretraining help?: Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law. ICAIL, pp 159–168. https://doi.org/10.1145/3462757.3466088

Zhou Y, Qin Y, Huang R, Chen Y, Lin C, Zhou Y (2024) Self-training improves few-shot learning in legal artificial intelligence tasks. Artif Intell Law. https://doi.org/10.1007/s10506-024-09403-z

Clause definition references

Boessel B (n.d.) Are anti-assignment clauses Enforceable?. Kira Systems. https://kirasystems.com/learn/are-anti-assignment-clauses-enforceable/. Accessed 16 Jul 2024

Non-Transferable Licenses – Definition (n.d.) Law Insider. https://www.lawinsider.com/dictionary/non-transferable-licenses. Accessed 16 Jul 2024

Sneddon (n.d.) Quickguide limitation and exclusion of liability. Ashurst. https://www.ashurst.com/en/insights/quickguide-limitation-and-exclusion-of-liability/. Accessed 17 Jul 2024

Parts of a Contract: Everything You Need to Know (2020) Upcounsel. https://www.upcounsel.com/parts-of-a-contract. Accessed 17 Jul 2024

Commercial and Technological Contracts Mythbuster - Backdating a Contract (2020) Stevens & Bolton. https://www.stevens-bolton.com/site/insights/articles/commercial-technology-contracts-mythbuster-backdating-a-contract. Accessed 17 Jul 2024

General Contract Clauses: Term and Termination (n.d.) Practical Law. https://uk.practicallaw.thomsonreuters.com/2-507-0812?contextData=(sc.Default)&transitionType=Default&view=hidealldraftingnotes. Accessed 16 Jul 2024

Using Contractual Risk Allocation Provisions to Minimize Risk and Maximize Reward (n.d.) Practical Law. https://uk.practicallaw.thomsonreuters.com/5-532-2743?contextData=(sc.Default)&transitionType=Default&firstPage=true. Accessed 16 Jul 2024

Miller (2022) Contracts and IP ownership. Thomson Reuters. https://legal.thomsonreuters.com/en/insights/articles/contracts-and-intellectual-property-ownership. Accessed 08 Jul 2024

Source Code Escrow (n.d.) Thomson reuters practical law. https://uk.practicallaw.thomsonreuters.com/6-502-4093?contextData=/28sc.Default%29&transitionType=Default Accessed 08 Jul 2024

Rishi A (2020b) The complete list of standard clauses to check before signing a contract. Spotdraft. https://www.spotdraft.com/blog/standard-clauses-to-check-contract. Accessed 17 Jul 2024

Non-Disparagement Provision (n.d.) Practical law. https://uk.practicallaw.thomsonreuters.com/9-584-2826?contextData=(sc.Default)&transitionType=Default&firstPage=true. Accessed 17 Jul 2024

The Noncompete Clause Explained (2021) Contracts counse. https://www.contractscounsel.com/b/noncompete-clause. Accessed 16 Jul 2024

Non-Solicitation (n.d.) Practical Law. https://uk.practicallaw.thomsonreuters.com/7-382-3652?contextData=%28sc.Default%29&transitionType=Default. Accessed 17 Jul 2024

What are Perpetual, Irrevocable, Royalty-free Licences (n.d.) Lexis Nexis. https://www.lexisnexis.co.uk/legal/guidance/what-are-perpetual-irrevocable-royalty-free-licences. Accessed 17 Jul 2024

Licensor vs Licensee - What's the Difference? (n.d.) WikiDiff. https://wikidiff.com/licensor/licensee. Accessed 16 Jul 2024

Guidance on Licensing Intellectual Property (n.d.) UK Government - Gov.uk. https://www.gov.uk/guidance/licensing-intellectual-property. Accessed 17 Jun 2024

What does Limited or Unlimited Mean in a License? (2017) Stack exchange. https://law.stackexchange.com/questions/24130/what-does-limited-or-unlimited-mean-in-a-license). Accessed 16 Jul 2024

General Contract Clauses: Audit Rights (n.d.) Practical law. https://uk.practicallaw.thomsonreuters.com/4-567-1110?contextData=(sc.Default)&transitionType=Default&view=hidealldraftingnotes. Accessed 17 Jul 2024

Change of Control Clause (n.d.) Practical law. https://uk.practicallaw.thomsonreuters.com/0-382-3325?contextData=/28sc.Default/29&transitionType=Default. Accessed 17 Jul 2024

Kagan J (2022) What is a covenant not to sue?. Investopedia. https://www.investopedia.com/terms/c/covenant-not-sue.asp. Accessed 16 Jul 2024

Exclusivity Agreement (n.d.) Practical law. https://uk.practicallaw.thomsonreuters.com/4-107-6577?contextData=(sc.Default)&transitionType=Default&firstPage=true. Accessed 16 Jul 2024

Governing Law (n.d.) Practical law. https://uk.practicallaw.thomsonreuters.com/8-107-3850?contextData=(sc.Default)&transitionType=Default&view=hidealldraftingnotes. Accessed 15 Jul 2024

Insurance Clause (n.d.) Contracts counsel. https://www.contractscounsel.com/g/41/us/insurance-clause. Accessed 15 Jul 2024

Gledhill L (2022) Joint Ownership of Intellectual property rights. Harper James. https://harperjames.co.uk/article/joint-ownership-of-ip/. Accessed 15 Jul 2024

Kenton W (2022) Most-favored nations (MFN) clause: treating other people equally. Investopedia. https://www.investopedia.com/terms/m/mostfavorednation.asp. Accessed 15 Jul 2024

Consequences of Termination Clause - Post-Termination Rights (n.d.) Hall Ellis Solicitors. https://hallellis.co.uk/consequences-termination-clause/. Accessed 16 Jul 2024

Price Restrictions (n.d.) Law Insider. https://www.lawinsider.com/clause/price-restrictions. Accessed 15 Jul 2024

Revenue Sharing Agreement (n.d.) Contracts Counsel. https://www.contractscounsel.com/t/us/revenue-sharing-agreement. Accessed 17 Jul 2024

Shareholder's Agreements: Right of First Refusal Versus Right of First Offer (2017) Mondaq. https://www.mondaq.co.uk/canada/shareholders/646438/shareholders39-agreements-right-of-first-refusal-versus-right-of-first-offer. Accessed 15 Jul 2024

Right of First Negotiation, Offer, and Refusal (n.d.) Practical law. https://uk.practicallaw.thomsonreuters.com/6-534-6258?contextData=/28sc.Default/29&transitionType=Default#co_anchor_a832508. Accessed 15 Jul 2024

General Contract Clauses: Third-Party Beneficiaries (n.d.) Practical law. https://uk.practicallaw.thomsonreuters.com/6-519-7630?contextData=(sc.Default)&transitionType=Default&view=hidealldraftingnotes. Accessed 15 Jul 2024