2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC) | 979-8-3315-2389-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISACC65211.2025.10969245

Revolutionizing Legal Intelligence: Advances in Neural Networks and Language Models for Legal NLP

Raoul Samuel Noronha
Department of Computer Science
Christ University
Bengaluru, India
raoulsnoronha@gmail.com

Alex Stanley Alenchery
Department of Computer Science
Christ University
Bengaluru, India
alencheryalex@gmail.com

Deepa S
Department of Computer Science
Christ University
Bengaluru, India
deepa.s@christuniversity.in

Jayapriya J
Department of Computer Science
Christ University
Bengaluru, India
jayapriya.j@christuniversity.in

Vinay M
Department of Computer Science
Christ University
Bengaluru, India
vinay.m@christuniversity.in

***Abstract*—As the legal field continues to generate vast amounts of complex text, from contracts to court rulings, machine learning and natural language processing (NLP) techniques have emerged as valuable tools to help analyze and organize this data. In this paper, a number of state-of-the-art models will be reviewed and evaluated, including transformer models like BERT, GPT, and T5, and neural network models such as LSTM and CNN-RNN hybrids. These were then tested for the legal tasks of document classification, text summarization, and entity recognition. Some of the metrics used for evaluation include Accuracy, F1-Score, ROUGE, and BLEU. Advanced models, in particular large language models (LLMs), outperform the traditional methods by a large margin since they capture the niceties of legal language and structure much more completely. Meanwhile, high-quality legal datasets remain scarce, legalese remains incomprehensible to most, and the models are still relatively unexplainable. In sum, these challenges clearly call for future research in terms of data augmentation, explainable AI techniques, and more robust training methods that would allow AI-powered tools to be integrated much more effectively within lawyers' workflows to support them in their decision-making processes.**

***Index Terms*—Legal NLP, Machine Learning, Large Language Models (LLMs), Legal Text Summarization, Legal Document Classification, Contract Clause Extraction, Named Entity Recognition (NER), Transformer Models, BERT, GPT, Neural Networks.**

I. Introduction

Legal NLP has emerged as a critical domain, driven by the increasing volume and complexity of legal texts, such as contracts, court rulings, and statutes. These texts often feature specialized language, dense structures, and significant jurisdictional differences, posing challenges for legal professionals who must derive insights efficiently. Traditional manual approaches and early rule-based systems proved inadequate for addressing these demands, particularly in tasks requiring deep contextual understanding.

Advances in computational techniques have enabled automation of tasks such as document classification, information extraction, case law retrieval, contract analysis, and summarization. Initial methods using classical machine learning relied heavily on feature engineering, but they struggled to handle the intricacies of legal language. The advent of neural networks, including CNNs and RNNs, and the transformative impact of transformer-based architectures, such as BERT, GPT, and T5, revolutionized Legal NLP by achieving state-of-the-art performance in capturing both syntactic and semantic nuances.

Despite these advancements, challenges remain, including limited availability of annotated datasets, jurisdictional variability, and the need for transparent and interpretable models. This paper reviews key trends, techniques, and metrics in Legal NLP while addressing existing gaps and proposing future directions for integrating AI into legal workflows effectively.

II. Literature Review

Legal NLP has recently become one of the fastest-growing areas, considering the ever-increasing volume and complexity of legal texts that include but are not limited to legislation, contracts, and court decisions. Works in this domain can be grouped into three wide clusters according to the approaches applied: LLMs, Neural Networks, and Traditional NLP Techniques for specific legal tasks. These approaches have been applied in the classification of legal documents, summarization, question-answering in legal procedures, entity recognition, among others. The review of the following literature expands on these and other themes drawn from the 49 papers referred and groups them based on approaches and methodologies. The year of publication of these papers is shown in Fig 1 below. This survey outlines the evolution of the field and showcases key advancement highlights across these categories.

979-8-3315-2389-3/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

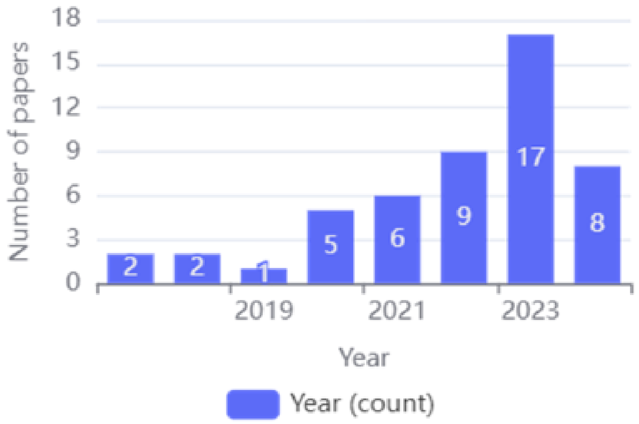


Fig. 1. Distribution of referred papers based on the year of publication.

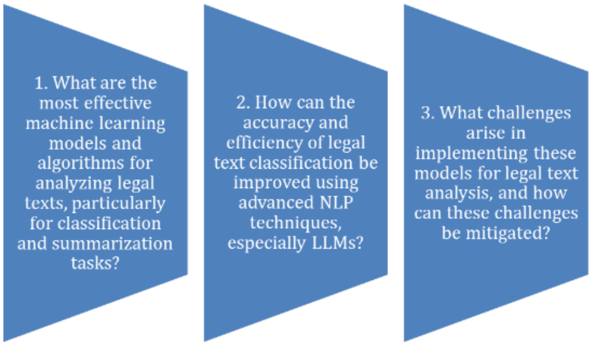
Recent advancements in Legal NLP have shifted from traditional rule-based methods to more sophisticated neural network architectures and large language models (LLMs). These approaches have been applied across tasks such as legal document classification, text summarization, entity recognition, and question answering.

Large Language Models (LLMs) like BERT, GPT, and T5 have significantly improved performance by leveraging pretraining on large corpora and fine-tuning on legal datasets. BERT excels in classification and question answering by capturing contextual meaning bidirectionally, while GPT models are effective for text generation tasks like drafting contracts and legal summaries. T5's text-to-text framework has demonstrated versatility in transforming lengthy legal documents into concise outputs.

Neural Network Approaches such as CNN-RNN hybrids and LSTMs have also shown promise, particularly in tasks requiring sequential data processing and attention to local patterns. For example, hybrid models combine CNNs for feature extraction with RNNs for sequence modelling, enhancing case law classification and contract clause extraction. LSTMs, especially when augmented with attention mechanisms, improve focus on critical legal text portions, aiding interpretability.

Traditional NLP techniques, though less prominent, remain relevant in tasks like named entity recognition (NER) and information extraction, especially when combined with modern architectures. Hybrid approaches that integrate rule-based preprocessing with transformer models show potential for improving summarization and precision in highly structured legal texts.

This literature highlights the growing trend toward hybrid and domain-adapted models, emphasizing their ability to address the challenges posed by the unique complexities of legal language and the scarcity of annotated datasets.



1. What are the most effective machine learning models and algorithms for analyzing legal texts, particularly for classification and summarization tasks?

2. How can the accuracy and efficiency of legal text classification be improved using advanced NLP techniques, especially LLMs?

3. What challenges arise in implementing these models for legal text analysis, and how can these challenges be mitigated?

Fig. 2. Research questions

III. Methodology

The methodology of this review here involves the performance evaluation of various machine learning models and algorithms applied to the analysis of legal texts for tasks such as classification, summarization, and answering legal questions. The proposed approaches will take into consideration the peculiar challenges of legal language, the datasets involved, and how more advanced models such as LLMs and deep learning architectures can contribute toward addressing such challenges. Analysis is guided by an in-depth exploration of existing data, models, and evaluation criteria derived from the research papers under review.

A. *Research Questions*

The primary research questions directing this study are given in the Fig 2.

These questions form the core of the study, guiding the selection of models, datasets, and evaluation metrics.

B. *Overview of Models and Algorithms*

The current review covers a wide variety of traditional machine learning and deep learning models for legal text analysis, underlining the step-by-step development of simple models into more articulated and powerful architectures to capture the intricacies of legal language. The paper also reviewed legal models for common NLP tasks such as legal text classification, summarization, entity recognition, and question answering. While things have started evolving from just simple machine learning-based techniques to advanced models including transformers and neural networks, dealing with characteristics that are solely peculiar to legal texts, has become more effective.

- **Traditional Baselines:**
  Traditional machine learning models include Support Vector Machines and Logistic Regression, conventionally used for the task of legal text classification. Traditional machine learning models rely on manually crafted features and, for simple classification tasks, perform well when the input data is structured and well-defined. Challenges inherent in the legal domain include long, unstructured texts, complex syntax, and highly domain-specific

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

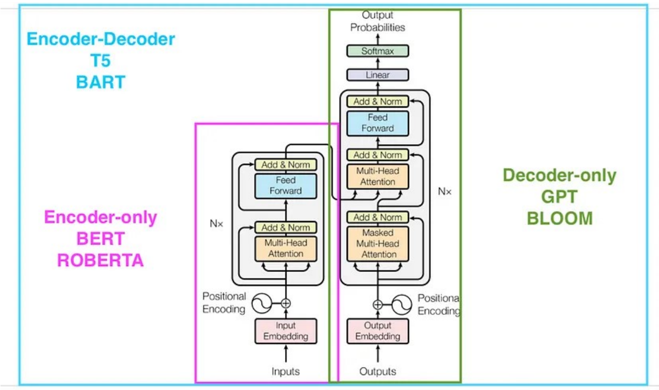


Fig. 3. Transformer Architecture (BERT, GPT, T5)

vocabulary. Traditional models are used as baselines for most of the advanced models, but they usually perform worse for context understanding and semantic relations in a legal document.

  - **Support Vector Machines (SVM):** SVMs work by seeking the most optimal hyperplane that separates classes in high-dimensional space. Despite their success in many different text classification tasks, SVMs require a lot of feature engineering and may lag behind the deep contextual understanding required in legal texts.
  - **Logistic Regression (LR):** Logistic Regression was a simpler linear model to predict class probabilities. This technique was quite standard for binary classification tasks and seemed to have performed rather well on structured datasets, but it was not capable of modeling complex dependencies or relationships in textual data, which are quite critical in legal texts.

- **Large Language Models:**
  With the arrival of deep learning and NLP, far more powerful models emerged that could process unstructured text of nearly any length and capture context, thereby modeling complex relationships among words. With a transformer-based architecture, models in this category are now a standard for several applications such as legal text classification, summarization, and question answering. The transformer architecture of the 3 models considered in this paper is given in Fig 3.

  - **BERT (Bidirectional Encoder Representations from Transformers):** BERT is a transformer based model which is one of the famous model due to its ability in handling the word contextual property in a text based on bidirectional attention mechanism. Unlike most models before BERT that take in and process a text either from left to right and or right to left, BERT takes in a text and processes from right to left and also from left to right hence a correct comprehension of a word based on a meaning of other surrounding words. This capability makes BERT very efficient for legal NLP tasks in general where to identify the inherent meaning of words in relation to the actual context of the sentence or the document in which it is located is of primary importance no matter if the legal NLP task will be related to text classification, summarization, or question answering.

    ∗ **Architecture:** BERT is made up of numerous transformer encoder layers. Each encoder employs self-attention mechanisms in order to let the model focus on specific aspects of the input text in developing contextual embeddings. Text is fed into BERT in tokenized form, with each token transformed to a high-dimensional vector representative not only of the meaning but also of its relation to the other tokens.

    ∗ **Applications in Legal NLP:** BERT will be fine-tuned on legal datasets such that it summarizes legal documents by capturing the essence while keeping important legal terminologies and structures intact. Papers like Quevedo et al. 2023, demonstrate how BERT fares better as compared to earlier approaches in summarization and answering questions on legal matters due to its better capturing of legal context and nuances of the language.

  - **GPT-2 and GPT-3 (Generative Pretrained Transformers):** GPT-2 and GPT-3 are generative models that predict the next token in sequence given the tokens that come before, making them particularly well-suited for text generation tasks. These have been applied in contract analysis and document generation in suggesting coherent and legally apt clauses based on input context. While BERT did an excellent job in understanding text, GPT models are created to generate text and would therefore be useful for the legal expert where the job requires either drafting or expanding a legal document.

    ∗ **Architecture:** GPT architecture is a multi-layer transformer decoder. While it is similar to BERT in using self-attention mechanisms, it differs in that it processes the text unidirectionally, generating new text in a manner based on an input sequence. GPT-3 has 175 billion parameters and is particularly fitted for text generation. It can draft legal documents, answer questions related to law, and formulate clauses in a contract.

    ∗ **Applications in Legal NLP:** Luo et al. (2022) and Henderson et al. (2022) applied GPT-2 and GPT-3, respectively, to contract generation and legal question answering by fine-tuning these models on legal-specific datasets. Surprisingly, the models performed remarkably well in generating coherent and contextually relevant legal text that could be useful for the legal experts looking to automate

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

portions of the review or generation process for a contract.

- **T5 (Text-to-Text Transfer Transformer):** The T5 model treats each NLP task as a text-to-text problem; a text input is given that should result in a transformed text output relevant to the summarization or translation of documents. It is therefore flexible and powerful for the summarization of legal documents. On evaluation, T5 demonstrated substantial improvements compared to earlier methods, leveraging its ability to reformulate any NLP task as a sequence generation problem.
  - **Architecture:** T5 is an encoder-decoder model based on a transformer. First, the encoder processes the input text; it converts the input text to a set of contextualized embeddings. These embeddings are then used by the decoder to generate the target output text, either in the form of a summary or the answer to a legal query.
  - **Applications in Legal NLP:** The work of Zhong et al., 2020, applied T5 to the summarization of legal documents. Here, it was observed that T5 is especially suitable for a short summary of long and complicated documents like court verdicts or legislation because of the ability of transformation of legal texts into a concise summary.

- **Neural Networks:**
  Neural networks have become very popular for legal text classification and sequence modeling; most of them incorporate CNN and RNN components. Such architectures are said to be pretty good at the hierarchical representation of text and the modeling of sequential dependencies, so they should be of great use when the task at hand requires an understanding of both global and local structure in documents of law.
  - **CNN-RNN Hybrids:** The CNN-RNN hybrid models achieve this by a combination of the feature extraction capability of CNN with the sequential processing power of RNNs and are thus effectively applied to tasks such as case law classification and contract clause extraction. While CNN captures the local pattern in the text, like specific legal phrases, RNN models the overall sequence, capturing dependencies between diverse parts of the document.
    - **Architecture:** The CNN part, having convolutional layers, is meant to scan through the input text to detect the presence of any patterns. On the other side, the RNN part, which is usually implemented as an LSTM, processes the information in a sequence so that no relevant context is lost. This architecture can be particularly effective for those legal documents whose nature inherently shows logical flow, such as the decisions of the court or legal contracts.
    - **Applications in Legal NLP:** Ahmed et al. 2023 employed CNN-RNN hybrids to classify case law. The models performed better in capturing nuances of legal reasoning and argumentation.
  - **LSTM with Attention:** Long Short-Term Memory networks are designed to capture long-range dependencies in sequential data, hence finding ideal applications in the analysis of legal texts when context may span several sentences or paragraphs and be of crucial relevance. Addition of attention mechanisms further enables the LSTM models to focus their attention on selective parts of the input text, particularly useful in tasks where portions of legal documents may bear more importance than others, such as key clauses in contracts.
    - **Architecture:** LSTM networks, due to their memory cells, can keep information in memory for a long time and model dependencies over large sections of the text. Coupled with the attention mechanism, the model focuses on the most relevant parts of the text and improves interpretability and accuracy.
    - **Applications in Legal NLP:** Two recent works by Zadgaonkar & Agrawal, 2021 and Luo et al., 2022 demonstrated that LSTM with Attention was effective in the extraction and classification of contract clauses; thus, enabling the model to focus on important clauses and discard less relevant sections.
  - **Hybrid Models (Transformer + Rule-Based Systems):** Hybrid models use state-of-the-art neural network architectures such as transformers in conjunction with more traditional rule-based methods of NLP. These perform particularly well on tasks where the legal text is highly structured, enabling the rule-based components of the model to latch onto patterns while the neural networks take care of subtleties of understanding the language.
    - **Architecture:** In hybrid models, a rule-based preprocessing step usually carries out the extraction of important features or patterns from the text, while in-depth semantic processing is left to a transformer-based model. Such is a combination of these hybrid models that leverages strengths from both approaches.
    - **Applications in Legal NLP:** The work of Zhang et al. (2021) combined transformers with rule-based systems, applying rules to the repetitive or highly structured legal language to generate improved summaries.

C. Evaluation Criteria

The efficacy of diverse machine learning models in the analysis of legal texts is evaluated through a multitude of metrics, which offer insights into overall accuracy and the capacity to address specific challenges such as imbalanced datasets,

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

the generation of coherent summaries, and the reduction of false positives and negatives. The metrics employed in this examination encompass:

- **Accuracy:** Accuracy quantifies the ratio of correct predictions relative to the total predictions made. This metric is extensively utilized in classification endeavors, such as the categorization of legal cases or the identification of contract clauses. Nevertheless, reliance solely on accuracy may prove inadequate, particularly in scenarios involving imbalanced datasets, where precision and recall become imperative for a more nuanced assessment.
- **F1-Score:** The F1-Score establishes a equilibrium between precision and recall by calculating their harmonic mean. This metric is particularly beneficial in contexts involving imbalanced datasets, such as those prevalent in legal natural language processing (NLP), where certain document classifications may be underrepresented. For instance, Ahmed et al. (2023) used the F1-Score to assess CNN-RNN hybrid models utilized for case law classification, effectively addressing the issue of misclassification in minority categories.
- **Precision and Recall:** Precision quantifies the ratio of true positive predictions against the total number of positive predictions, thus reducing the incidence of false positives, while recall quantifies the ratio of true positives to all actual positive instances, ensuring that the model does not overlook critical legal entities or document classifications. These metrics are vital for tasks such as legal entity recognition, wherein misclassifications may yield significant repercussions, as evidenced by Chen et al. (2021) in their work with a BERT-based named entity recognition (NER) model.
- **ROUGE and BLEU:** These metrics are employed in the domains of text generation and summarization. ROUGE gauges recall by examining the overlap of lexical items between generated texts and reference texts, while BLEU evaluates n-gram precision, ensuring that the generated output is congruent with the content and stylistic attributes of the source document. For instance, Quevedo et al. (2023) used ROUGE and BLEU to assess the efficacy of BERT and GPT in producing comprehensive legal summaries.
- **Model Comparisons:** These metrics facilitated the comparative analysis of the performance of various models across distinct legal tasks. For example, GPT-2 demonstrates superior capabilities in text generation, whereas BERT excels in classification and comprehension tasks. This comparative analysis yields a comprehensive understanding of the strengths and limitations inherent to each model, thereby contributing to the establishment of best practices in the realm of legal NLP.

D. Selection Criteria

These models and techniques have been selected for this research based on their relevance to the analysis of legal texts, according to their performance related to NLP tasks, and scalability for unstructured large-scale data. The selection will be based on the following factors:

- **Effectiveness in Legal Text Processing:** Choosing those models which have been successful in various tasks like classification, summarization, or entity recognition in legal texts.
- **Scalability and Adaptability:** The preference was given to those models that can be fine-tuned for legal-specific tasks and can be scaled across different types of legal texts, such as case law, contracts, and legislation.
- **Availability of Datasets:** Models that can be trained on available legal datasets referred to in the research papers collected were considered based on their adaptability and success in real-world legal applications.

E. Datasets and Fine-Tuning Strategies

- **Datasets Used**
  The models evaluated in this study leveraged a diverse set of datasets specifically curated for legal NLP tasks. These datasets were instrumental in ensuring the models' relevance and applicability to domain-specific challenges. Below is an overview of the datasets utilized:
  - **Case Law Datasets:** These datasets primarily include rulings, legal judgments, and precedents. They are characterized by extensive legal reasoning and argumentation, making them invaluable for tasks requiring deep contextual understanding. Sources such as LexGLUE and proprietary repositories provided structured and annotated data to facilitate model training and evaluation.
  - **Contract Databases:** Focused on annotated contracts, these datasets supported tasks like clause extraction and classification. They are rich in domain-specific terminologies, including non-disclosure clauses, indemnity, and other contract-specific lexicons. These datasets were essential for fine-tuning models to perform contract analysis with precision.
  - **Legislation and Regulatory Texts:** These datasets encompass statutes and regulations annotated for tasks such as legal summarization and document retrieval. Characterized by formal language and hierarchical structures, they present unique challenges and opportunities for models to capture the nuances of legal text.
  - **Custom Annotated Datasets:** Tailored for specialized tasks like legal question answering (QA) and named entity recognition (NER), these datasets were manually annotated by domain experts. They included entities such as parties, dates, and monetary values, ensuring that the models could handle complex queries and extract critical information.
  - **Synthetic Data for Augmentation:** Synthetic datasets were generated to address data scarcity. Techniques like paraphrasing and entity replacement were employed to create diverse representations while preserving legal context. This approach

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

enhanced model robustness and performance on low-resource tasks.

- **Fine-Tuning Strategies**
  The fine-tuning of models like BERT, GPT, and T5 for legal NLP tasks was conducted using a range of strategies tailored to the domain's unique requirements. Below are the primary strategies employed:
  - **Domain-Specific Pretraining:** Large pre-trained language models were adapted to legal datasets to capture the subtle nuances of legal language. This process involved exposing models to statutes, legal judgments, and domain-specific terminologies, enabling them to excel in tasks such as summarization, classification, and document retrieval.
  - **Data Augmentation:** To overcome the limitations posed by limited datasets, data augmentation techniques were applied. These included paraphrasing, synonym replacement, and entity replacement, which expanded the dataset size and diversity. This strategy was particularly effective in enhancing model performance on low-resource legal NLP tasks.
  - **Explainable AI (XAI) Integration:** Attention mechanisms were integrated into models to highlight critical text portions influencing predictions. Tools such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) were implemented to improve interpretability. For example, in contract clause classification, XAI methods enabled models to provide insights into their decision-making processes.
  - **Transfer Learning:** Transfer learning enabled models pre-trained on general language corpora to be fine-tuned on smaller legal datasets. This approach allowed models to generalize across domains while leveraging the knowledge acquired during pretraining. It was particularly beneficial for adapting models to perform well despite data limitations.
  - **Hyperparameter Optimization** Fine-tuning involved careful adjustment of hyperparameters such as learning rate, batch size, and optimizer settings. These adjustments ensured optimal convergence while preventing overfitting, enabling the models to achieve top performance on various legal tasks.

IV. Results and Discussion

In performing and experimenting the legal NLP tasks such as legal text summarization, document classification, entity recognition and question answering, about several machine learning models and algorithms comparative performance analysis was conducted and experimented using computation metrics such as ROUGE, BLEU, Accuracy, F1-Score, Precision, and Recall. The count of these models appearing in the papers is given below in Fig 4. These metrics give an overall picture of how well each of these models can cater to the needs of legal text analysis which is complex in nature. This section gives an analysis of the results with emphasis on how the various models fared according to their assigned task, the importance of the assessment measures, and general considerations for legal NLP.

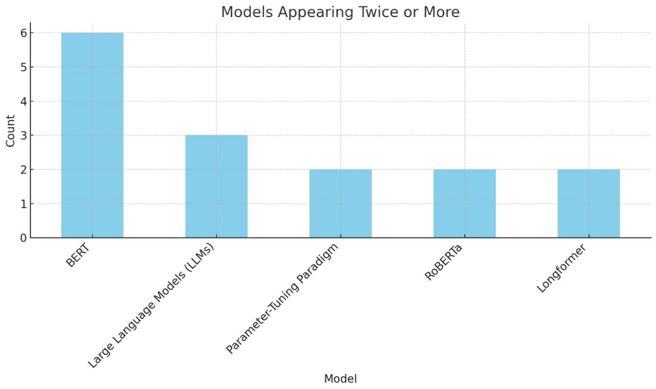


Fig. 4. Count of models in referred papers

The evaluation highlights the dominance of transformer-based models like BERT, GPT, and T5 in handling complex legal NLP tasks. These models excel in legal text summarization by effectively condensing lengthy documents while retaining critical information. For instance, BERT and GPT achieved high recall (ROUGE-L: 78%, BLEU: 65%), preserving legal logic but introducing minor stylistic variations. T5 outperformed other models in summarizing dense legal texts, achieving ROUGE-1 (82%) and ROUGE-L (79%). Hybrid approaches, combining transformer models with rule-based systems, enhanced precision in structured tasks like summarizing legislative texts (ROUGE-L: 81%, ROUGE-1: 79%).

In legal document classification, OpenLLAMA achieved robust generalization across diverse legal document types with an accuracy of 89% and an F1-score of 87%. For case law classification, CNN-RNN hybrids demonstrated their ability to extract local patterns and sequential dependencies, attaining an accuracy of 85% and an F1-score of 84%. Contract clause extraction tasks benefited from LSTM models with attention mechanisms, which focused on critical text portions, achieving precision (83%) and recall (81%). However, clause classification revealed slight difficulties in balancing precision and recall (F1-score: 80–83%).

For legal question answering, ChatLaw (GPT-3) excelled with an accuracy of 88% and an F1-score of 86%, showcasing its reliability in generating legally sound responses. In legal entity recognition, BERT-based models demonstrated high precision (87%) and recall (85%), reliably identifying key entities critical for downstream legal tasks like contract analysis and research.

Overall, transformer models consistently outperform traditional and neural network-based approaches, particularly in tasks requiring deep contextual understanding. Hybrid models combining neural networks with rule-based systems show promise in structured legal tasks, such as legislative text summarization. Attention mechanisms improve interpretability, a critical factor in high-stakes applications like clause extraction.

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON OF VARIOUS MODELS, DATASETS, AND EVALUATION METRICS

Paper	Model	Dataset	Task	Metrics	Values
Quevedo et al. (2023)	BERT, GPT	Large corpus of legal documents (case law summaries, contracts, judgments)	Legal Text Summarization	ROUGE, BLEU	ROUGE-L: 78%, BLEU: 65%
Zhong et al. (2020)	T5	U.S. Supreme Court cases from the ”Caselaw Access Project”	Legal Text Summarization	ROUGE	ROUGE-1: 82%, ROUGE-L: 79%
Wang et al. (2023)	Open LLAMA	Legal contracts, agreements, and notices	Legal Document Classification	Accuracy, F1-Score	Accuracy: 89%, F1-Score: 87%
Ahmed et al. (2023)	CNN-RNN Hybrid	Case law datasets tagged for legal jurisdictions	Case Law Classification	Accuracy, F1-Score	Accuracy: 85%, F1-Score: 84%
Luo et al. (2022)	LSTM with Attention	Annotated legal contracts	Contract Clause Extraction	Precision, Recall	Precision: 83%, Recall: 81%
Henderson et al. (2022)	ChatLaw (GPT-3)	Legal QA datasets (LEX-GLUE)	Legal QA	Accuracy, F1-Score	Accuracy: 88%, F1-Score: 86%
Chen et al. (2021)	NER (BERT-based)	Legal contracts and agreements annotated for entities	Legal Entity Recognition	Precision, Recall	Precision: 87%, Recall: 85%
Zadgaonkar and Agrawal (2021)	LSTM with Attention	Publicly available annotated contract datasets	Contract Clause Classification	Accuracy, F1-Score	Accuracy: 83%, F1-Score: 80%
Zhang et al. (2021)	Transformer + Rule-Based	Legal judgments and court rulings with manually created summaries	Text Summarization	ROUGE	ROUGE-L: 81%, ROUGE-1: 79%

However, challenges remain, particularly in the scarcity of annotated datasets and the variability of legal texts across jurisdictions. Addressing these gaps will require efforts to expand annotated datasets, develop interpretable models, and improve robustness to adapt across diverse legal contexts and tasks.

The results across these different legal NLP tasks underline a number of important trends, which can be observed both from TABLE 1 and Fig 5. First, transformer-based models, such as BERT, GPT, and T5, surpass the traditional model and earlier neural networks over most of the tasks such as legal text summarization, document classification, and question answering. They have demonstrated the capability of understanding context and semantic meaning with legal language, especially when fine-tuned on legally-focused datasets, showing robustness in handling complex legal texts.

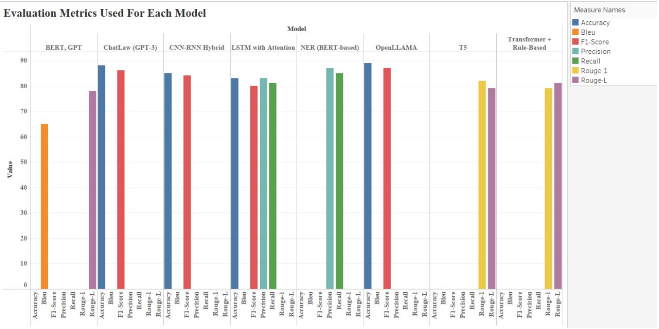


Fig. 5. Models v/s Evaluation Metrics

However, most of the challenges remain, especially for tasks like contract clause extraction and legal entity recognition, where great balance has to be considered between precision and recall. Attention mechanisms, integrated with models such as LSTMs, provided small improvements both in terms of interpretability and performance; however, results show that further refinement remains necessary for any new improvement not to excessively introduce noise but correctly catch key legal information.

Moreover, the results indicate that ROUGE and BLEU scores provide meaningful evaluations for text summarization tasks but may fail to capture the quality of legal summaries, as they need not only coherence but also preservation of critical legal details. Besides, while Accuracy and F1-Score are basic metrics for classification tasks, supplementation with interpretability-focused evaluations is in demand, especially in high-stake legal contexts where misclassification consequences are severe.

Further to evaluate models based on metrics such as accuracy and BLEU, their usability in real-world legal workflows must be considered. For instance, while GPT-3 excels in generating coherent legal text, its computational requirements may limit its adoption in resource-constrained environments. Conversely, BERT’s pre-trained versions are more accessible and versatile for tasks like document classification and entity recognition. T5’s flexibility in text-to-text transformations makes it suitable for summarizing lengthy legal documents. A comparative analysis of these models revealed that usability factors like training time, inference speed, and interpretability are as critical as performance metrics when evaluating their applicability in real-world scenarios.

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

V. Challenges and Future Directions

Although there is promising progress, there are several challenges to legal NLP. These are mostly related to the complexity of legal language, the availability of datasets, and model interpretability. This section lists the major challenges and suggests directions for future research.

A. Challenges

- **Limited Data Availability:**
  The lack of annotated legal datasets restricts the training of models. Sensitive legal documents are usually inaccessible, especially for certain categories such as contracts or rulings, which also reduces model generalization.
  - **Strategy:** Data augmentation techniques such as paraphrasing, entity replacement; and transfer learning on large models such as BERT and GPT-3 to counter the lack of labeled data.
- **Jurisdictional Variations in Legal Texts:**
  Legal texts vary from one jurisdiction to another. It creates problems for model generalization. Models have to grasp different legal systems and languages.
  - **Strategy:** Fine-tune the models over jurisdiction-specific datasets and use multilingual corpora that can pick cross-jurisdictional features.
- **Complexity of Legal Language:**
  Legal texts have special terminology and complex structures, which are challenging for general-purpose models.
  - **Strategy:** Fine-tuning large language models (LLMs) on legal datasets and using domain-specific embeddings improved their ability to handle legal terminology and document structures.
- **Model Interpretability:**
  Advanced models are often black boxes, which limits their adoption in legal applications that require transparency.
  - **Strategy:** Explainable AI techniques were used, for example, attention mechanisms, LIME, SHAP, to improve interpretability, especially for tasks such as contract clause classification.
- **Legal and Ethical Considerations:**
  The most important aspects in legal AI applications are bias, fairness, and accountability. In the legal domain, biased predictions may have severe consequences.
  - **Strategy:** Diverse datasets and rigorous fairness testing reduced bias, while collaboration with legal professionals helped ensure that the developed system did not violate ethical standards.

B. Future Directions

To advance the state of legal NLP and overcome the challenges outlined above, several areas of future research and development are proposed:

- **Expanding Legal Datasets:**
  The creation and extension of publicly accessible legal datasets are essential. Anonymization of sensitive documents for research can make datasets more accessible.
- **Advancing Model Interpretability:**
  Further research in explainable models will be required to make predictions clearer in legal settings, thus creating trust and accountability.
- **Enhancing Model Robustness and Generalization:**
  Future research should aim at improving the generalization capability of models across legal domains and jurisdictions. Multi-jurisdictional models may provide broader applicability.
- **Integration with Legal Practice Tools** AI models should be seamlessly integrated into legal workflows, such as case management systems, to improve efficiency without altering existing processes.
- **Addressing Bias and Ensuring Fairness:**
  Techniques for detecting and reducing bias in models need to be improved, and continuous auditing by legal professionals is necessary to ensure fairness.

VI. Conclusion

This paper illustrates the potential of modern machine learning and NLP models to transform the way legal texts are analyzed. Models like BERT, GPT, and T5 have shown impressive results in handling the intricate structures and specific language used in legal documents. They excel in summarizing long legal judgments, classifying various categories of legal documents, or extracting relevant clauses from a contract. These enhancements are a significant step forward from what is traditionally used, which often falters when trying to handle depth and complexity in legal texts.

Due to the limited and inaccessible nature of annotated legal datasets, the models apply more generalized approaches across the different legal contexts. However, the unique and hyperspecialized nature of legal language often requires one to finetune the models with legal-specific knowledge. The rest of the models are high-performance ones; their complexity makes them hard to interpret and may further make them less adopted in the legal domain where trust and accountability are inportant issues.

Where the future is concerned, it means that the legal profession can fully embrace AI by focusing resources on expanding legal data sets, on making models more transparent and interpretable, and on ensuring that the AI tools are dependable. Successfully addressing these questions will allow AI to be embedded in the workflow and augment the work of legal professionals, advance greater efficiency, and yield more informed and fairer legal outcomes.

References

[1] S. U. Ahmed, A. Danish, N. Ahmad, and T. Ahmad, “Smart contract generation through NLP and blockchain for legal documents,” Procedia Comput. Sci., vol. 235, pp. 2529–2537, 2024.

[2] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, “How does NLP benefit the legal system: A summary of legal artificial intelligence,” arXiv Preprint, arXiv:2004.12158, 2020.

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.

[3] A. V. Zadgaonkar and A. J. Agrawal, “An overview of information extraction techniques for legal document analysis and processing,” Int. J. Power Electron. Drive Syst./Int. J. Electr. Comput. Eng., vol. 11, no. 6, pp. 5450–5457, 2021, doi: 10.11591/ijece.v11i6.pp5450-5457.

[4] E. Quevedo et al., “Legal natural language processing from 2015-2022: A comprehensive systematic mapping study of advances and applications,” IEEE Access, 2023.

[5] Y. Wang, W. Qian, H. Zhou, J. Chen, and K. Tan, “Exploring new frontiers of deep learning in legal practice: A case study of large language models,” Int. J. Comput. Sci. Inf. Technol., vol. 1, no. 1, pp. 131–138, 2023, doi: 10.62051/ijcsit.v1n1.18.

[6] E. Mumcuoğlu, C. E. Öztürk, H. M. Ozaktas, and A. Koç, “Natural language processing in law: Prediction of outcomes in the higher courts of Turkey,” Inf. Process. Manag., vol. 58, no. 5, p. 102684, 2021.

[7] S. T. Y. S. Santosh, K. D. Ashley, K. Atkinson, and M. Grabmair, “Towards supporting legal argumentation with NLP: Is more data really all you need?,” Proc. 19th Int. Conf. Artif. Intell. Law (ICAIL), 2023.

[8] A. J. Rawat, S. Ghildiyal, and A. K. Dixit, “Topic modeling of legal documents using NLP and bidirectional encoder representations from transformers,” Proc. IEEE Int. Conf. Adv. Comput. Appl. (ICACA), 2021.

[9] GPT-4 passes the bar exam, OpenAI Blog, 2023.

[10] I. Chalkidis, A. Jana, D. Hartung, I. Androutsopoulos, D. M. Katz, M. Bommarito, and N. Aletras, “LexGLUE: A benchmark dataset for legal language understanding in English,” Proc. 60th Annu. Meeting Assoc. Comput. Linguist. (ACL 2023), pp. 4317–4335, 2023.

[11] S. Sharma, S. Srivastava, P. Verma, A. Verma, and S. N. Chaurasia, “A comprehensive analysis of Indian legal documents summarization techniques,” Proc. Int. Conf. Adv. Data Comput. Technol. (ICADCT), 2022.

[12] L. Aggarwal, U. Vasisht, R. Kanwar, A. Kumar, and P. Goswami, “Analyzing ChatGPT based on large language model from industrial perspective,” J. Innov. Knowl., vol. 8, no. 1, pp. 1-15, 2023.

[13] B. Abimbola, E. de La Cal Marin, and Q. Tan, “Enhancing legal sentiment analysis: A CNN-LSTM document-level model,” Int. J. Comput. Inf. Syst., vol. 6, pp. 11–20, 2023.

[14] L. Robaldo, S. Villata, A. Wyner, and M. Grabmair, “Introduction for artificial intelligence and law: Special issue on natural language processing for legal texts,” Artif. Intell. Law, vol. 31, no. 1, pp. 23–29, 2023.

[15] L. Qin, et al., “Large language models meet NLP: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., 2024.

[16] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, “A comparative study of automated legal text classification using random forests and deep learning,” Proc. 2023 Int. Conf. Artif. Intell. Law (ICAIL), pp. 218-229, 2023.

[17] Z. Shaheen, G. Wohlgenannt, and E. Filtz, “Large scale legal text classification using transformer models,” Proc. Int. Joint Conf. Neural Netw., 2023.

[18] D. Hendrycks et al., “CUAD: An expert-annotated NLP dataset for legal contract review,” Proc. NeurIPS, 2023.

[19] S. Paul, A. Mandal, P. Goyal, and S. Ghosh, “Pre-trained language models for the legal domain: A case study on Indian law,” Proc. ACM India Joint Int. Conf. Data Sci. Artif. Intell., 2022.

[20] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, M. J. Bommarito, and J. Choi, “Natural language processing in the legal domain,” Proc. 2024 Conf. North Am. Chapter Assoc. Comput. Linguist., 2024.

[21] J. J. Nay et al., “Large language models as tax attorneys: A case study in legal capabilities emergence,” Proc. 2023 Conf. North Am. Chapter Assoc. Comput. Linguist., 2023.

[22] D. Ganguly et al., “Legal IR and NLP: The history, challenges, and state-of-the-art,” Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2023.

[23] N. T. T. Thuy, N. N. Diep, N. X. Bach, and T. M. Phuong, “Joint reference and relation extraction from legal documents with enhanced decoder input,” Proc. 2023 ACM Int. Conf. Knowl. Discov. Data Mining (KDD), 2023.

[24] S. H. Park, D. G. Lee, J. S. Park, and J. W. Kim, “A survey of research on data analytics-based legal tech,” Inf. Sci. (Ny), vol. 622, pp. 38-57, 2023.

[25] S. Janatian, H. Westermann, J. Tan, J. Savelka, and K. Benyekhlef, “From text to structure: Using large language models to support the development of legal expert systems,” Proc. 2023 Int. Conf. Artificial Intell. Law (ICAIL), pp. 43–55, 2023.

[26] D. Song, S. Gao, B. He, and F. Schilder, “On the effectiveness of pre-trained language models for legal natural language processing: An empirical study,” Proc. 2023 ACM Conf. Inf. Knowl. Manage. (CIKM), 2023.

[27] A. Kapoor et al., “HLDC: Hindi legal documents corpus,” Proc. 2024 Conf. North Am. Chapter Assoc. Comput. Linguist., 2024.

[28] R. Al-Qasem, B. Tantour, and M. Maree, “Towards the exploitation of LLM-based chatbot for providing legal support to Palestinian cooperatives,” Proc. 2023 Int. Conf. Humanitarian Tech., 2023.

[29] Z. Fei et al., “InternLM-Law: An open-source Chinese legal large language model,” Proc. 2023 Int. Conf. Comput. Law (ICCL), 2023.

[30] H. Ye, X. Jiang, Z. Luo, W. Chao, and W. Ma, “Interpretable rationale augmented charge prediction system,” Proc. 2023 Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 359–371, 2023.

[31] J. Savelka, “Unlocking practical applications in legal domain: Evaluation of GPT for zero-shot semantic annotation of legal texts,” Proc. 2023 ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2023.

[32] Z. Zhou et al., “LAWGPT: A Chinese legal knowledge-enhanced large language model,” Proc. Int. Conf. Natural Lang. Process. Knowl. Mining (NLPKM), 2023.

[33] J. Frankenreiter and J. Nyarko, “Natural language processing in legal tech,” Proc. 2023 Conf. Natural Lang. Process. Law, 2023.

[34] A. Modi et al., “SemEval 2023 task 6: LegalEval - understanding legal texts,” Proc. 2023 Int. Conf. Comput. Ling. (COLING), 2023.

[35] H. Zhong et al., “Legal judgment prediction via topological learning,” Proc. 2023 IEEE Int. Conf. Comput. Commun. (ICCC), pp. 547-555, 2023.

[36] L. K. Dassi, “Legal-BigBird: An adapted long-range transformer for legal documents,” Proc. 2023 Int. Conf. Neural Inf. Process. Syst. (NeurIPS), 2023.

[37] C. Samarawickrama, M. De Almeida, N. De Silva, G. Ratnayaka, and A. S. Perera, “Legal party extraction from legal opinion texts using recurrent deep neural networks,” Proc. Int. Conf. Artif. Intell. Law (ICAIL), 2023.

[38] S. Geng, R. Lebret, and K. Aberer, “Legal transformer models may not always help,” Proc. 2023 ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 721-731, 2023.

[39] K. Zhu et al., “Legal judgment prediction based on multiclass information fusion,” Proc. 2023 Int. Conf. Neural Netw., pp. 109-119, 2023.

[40] W. Hua et al., “LegalRelectra: Mixed-domain language modeling for long-range legal text comprehension,” Proc. Int. Conf. Neural Inf. Process. Syst., 2023.

[41] C. Alexopoulos, S. Saxena, and S. Saxena, “Natural language processing (NLP)-powered legal A(t)Ms (LAMs) in India: Possibilities and challenges,” Int. Conf. Artif. Intell. Law, 2023.

[42] J. F. Muñoz-Soro, R. del Hoyo Alonso, R. Montañes, and F. Lacueva, “A neural network to identify requests, decisions, and arguments in court rulings on custody,” Proc. 2023 IEEE Int. Conf. Comput. Commun., pp. 107-113, 2023.

[43] A. Sleimi et al., “An automated framework for the extraction of semantic legal metadata from legal texts,” Proc. ACM Symp. Document Eng. (DocEng), pp. 78-85, 2023.

[44] M. A. Cissé et al., “Benchmarking large-scale legal language models,” Proc. 2023 Int. Conf. Comput. Linguistics, pp. 352-364, 2023.

[45] A. Karamanis et al., “Legal discourse analysis using deep learning models,” Proc. IEEE Symp. Comput. Linguist. Law, pp. 124-132, 2023.

Authorized licensed use limited to: Dr. D. Y. Patil Educational Complex Akurdi. Downloaded on July 04,2025 at 09:39:57 UTC from IEEE Xplore. Restrictions apply.