Leveraging Large Language Models for Natural Language Processing Based Tasks in the Legal Domain: A Short Survey

Yavuz Balı[1,2(✉)] and Birol Çiloğlugil[3]

[1] Institute of Natural Sciences, Department of Computer Engineering, Ege University, 35100 Bornova, Izmir, Turkey
9122001744@ogrenci.ege.edu.tr
[2] Intelligent Application Department, Huawei R&D Center, Istanbul, Turkey
[3] Faculty of Engineering, Department of Computer Engineering, Ege University, 35100 Bornova, Izmir, Turkey
birol.ciloglugil@ege.edu.tr

Abstract. The integration of natural language processing techniques (NLP) into the legal domain has opened up new possibilities for automating, analyzing, and improving access to legal information. However, the complex and specialized terminology of legal texts has always been a significant challenge for the NLP area. To overcome this challenge, research studies in the law domain conducted with NLP techniques can be categorized into the following tasks: classification, information extraction, summarization, information retrieval, legal judgment prediction. Recent advancements in large language models (LLMs) have significantly contributed to NLP tasks across various domains, including the aforementioned tasks in the legal domain. The ability of large language models to reason and understand the deep context of texts has shown promising capabilities for understanding legal language and overcoming legal challenges. Therefore, this paper explores the existing tasks addressed by NLP in the legal domain and delves into LLM-based approaches which address these NLP tasks. Moreover, the datasets and evaluation criteria for LLMs have been examined, and the challenges and key limitations of LLMs in the legal domain have been identified. Hallucination, lack of domain-specific knowledge, and lack of interpretability have been determined as the challenges for the reliable application of large language models in the legal field. Finally, opportunities for future research have been discussed.

Keywords: Law · Legal Domain · Large Language Models · Natural Language Processing

1 Introduction

It was asserted by [16] in 1987 that technology will standardize and automate a significant portion of legal services, and these advancements were defined as

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 166–178, 2025.
https://doi.org/10.1007/978-3-031-96997-3_11

Legal Tech. The processes within the legal domain are predominantly textual. Hence the domain is inherently a significant research area for Natural Language Processing (NLP). In this regard, there are many studies and approaches to facilitate legal processes. In particular with the development of artificial intelligence technologies, the volume of studies in Legal Tech is increasing considerably.

Rapidly enhancing artificial intelligence technologies affect many domain and change information processing approaches. These enhancements affect the legal domain as it affects other fields [19]. Hence, novel methods are being developed to address legal challenges. The volume of electronic legal documents and legal texts increases exponentially with digitization in the legal domain [1]. The types of documents encompass a broad range, including legal norms, contracts, court rulings, and patents. They are typically detailed and lengthy, with an inherently complex structure that makes exploration challenging [4]. As it depends on high volume of written text, law is an essential application field of NLP [23]. As a result, reasoning and comprehension can be difficult for legal documents [28]. In particular, the unique terminology of legal texts makes its significantly difficult to understand and analyze. The complexity and high amount of legal texts constitute crucial challenges even for the legal professionals [35].

One of the state-of-the-art techniques in the area of natural language processing, Large Language Models (LLMs), have significantly proven their superior performance in understanding natural language. In particular, the GPT-3 model developed by OpenAI attracts attention with its ability to produce human-like texts in diverse language tasks [6]. The models have made significant advancements in natural language processing by effectively learning the contextual and semantic features of language, thereby facilitating near-human performance in a wide range of language tasks [6]. One of the promising enhancements in the field of LLMs and their potential applications is the execution of legal tasks [33]. LLMs have demonstrated notable performance in various legal tasks [36]. Furthermore, the ability of Large Language Models to learn the contextual and semantic depths of the language carries great potential in understanding the difficult terminology of legal texts.

This paper deals with the examination of tasks addressing the difficulties in the area of law and the analysis of how recent studies leverage Large Language Models to address these tasks. We aim to overview the approaches to the challenges in the field of law with the emergence of Large Language Models. Thus, the contributions of this paper can be identified as listed below:

- An overview of the key Natural Language Processing (NLP) tasks that constitute the focus of research in the legal domain.
- An analysis of the suggested approaches leveraging Large Language Models for these tasks.
- An analysis of the data sources utilized by the approaches leveraging Large Language Models.
- An examination of the evaluation metrics employed by these approaches and the performance measures of Large Language Models.

The paper is organized follows: Sect. 2 reviews the tasks in the legal domain using natural language processing techniques, Sect. 3 delves into the novel approaches applied to these handle tasks by utilizing large language models, Sect. 4 presents the data resource used in the papers dealing with large language models, and Sect. 5 demonstrate an explanation of the evaluation strategies and evaluation metrics employed by LLMs, Sect. 6 provides a discussion about the limitations of large language models on natural language processing tasks in the legal domain, and Sects. 7 concludes the paper with final remarks.

2 Challenges and Tasks of the Legal Domain

The exponentially increasing volume of legal information, such as cases and legal documents, has encouraged many challenges in the legal domain. There are many studies on the challenges to help stakeholders in the legal domain. Particularly novel enhancements in the area of artificial intelligence have accelerated the efforts to tackle the challenges. One of these approaches is text classification. For lawyers and legal researchers, examining and researching big volume of legal texts, with new documents getting published constantly, is a time-consuming and challenging process. Vatsal et al. [34] states that classifying and searching US Supreme Court decisions published on the website requires a lot of manual effort by legal researchers. To address this challenge, they proposed automatic classification of court decisions by using artificial intelligent methods. They used the BERT [9], which is a transformer model. Moreover, in [23], the performance of machine learning and deep learning methods is evaluated for binary classification of case texts as "violation" and "no violation". They demonstrated that deep learning methods offer better performance than traditional statistical approaches. Shaheen et al. [27] also performed classification of legal documents according to their relevant areas by utilizing transformer models. In contrast to traditional classifications, they specifically addressed the challenge of large multi-label text classification.

On the other hand, several research studies have been conducted by [7,17, 20], to extract meaningful entities in the legal texts in order to automate the document analysis process of legal stakeholders. These studies leveraged the BiLSTM model, which is one of the deep learning techniques, for information extraction in legal documents and texts.

Working with legal texts emerge as another challenge due to their long-drawn and difficult to grasp structure. Therefore, text summarization is considered as another task to deal with these long-drawn and difficult to grasp texts. Considering the fact that the volume of electronic documents is increasing constantly, Banedetto et al. [4] recommended language-specific summarization and headline generation techniques to enhance accessibility of documents and make them easy to use. They also benefit from transformer models to tackle this challenge.

Lawyers and many legal researchers spend a significant amount of time searching for the most relevant legal document to support or refute a particular point of assertion. Also, this task is particularly challenging given the need for

high recall where high volume of legal documents is involved [31]. Accordingly, numerous research studies address the information retrieval as one of the main tasks in the legal domain. Thomas et al. [31] presented a legal research recommender system developed to suggest relevant precedents by extracting arguments from legal documents. The system aims to provide contextually meaningful decisions to the user by combining stages such as argument segmentation and case ranking. Tagarelli et al. [30] proposed a legal text mining approach that overlaps with retrieval-based tasks by automatically extracting and classifying legislative articles on the Italian civil code using the LamBERTa framework they developed.

As another task encountered in legal domain, Legal Judgment Prediction (LJP) focuses on predicting judgment results (including crimes, relevant articles, and length of sentence) from the fact description of a specific case [42]. Civil individuals may have difficulty finding answers to legal problems because they are not familiar with legal texts or provisions. Through LJP, legal support can be provided to civil individuals. Furthermore, in order to be able to provide references for research, numerous legal stakeholders could benefit from legal judgment prediction. For instance, in order to increase the interpretability of judgment prediction models, the proposed QAjudge model establishes the decision-making process in a human-like manner through question-answer structures [42]. The model selects relevant questions by taking into account legal principles, produces answers to them according to the case text, and makes a decision prediction based on these answers. A key strength of the approach is its ability to enhance the transparency and efficiency of the judicial process by reducing redundant questioning via reinforcement learning.

Considering the examined research studies, five distinct tasks emerge regarding the challenges in the legal area. These studies aim to tackle the tasks with different approaches and to open new directions to overcome the challenges. Table 1 presents the task categories and relevant research studies for each category based on the utilized approach as LLM based and non-LLM based.

3 Addressing Legal Challenges with Large Language Models

Recent studies have demonstrated that large language models possess strong reasoning capabilities and a deep understanding of linguistic and domain-specific semantics [6]. Given these abilities, LLMs hold significant potential in the legal domain, where the terminology is complex and precise language use is critical. Recently, LLMs have shown significant development and have been employed to a wide range of NLP downstream tasks across different domains [13]. For instance, the inherent verbosity and complexity of the legal documents may limit the accessibility of legal workers and practitioners to keep track of the updates [3]. Therefore, LLMs have great promise for the legal domain, since many researchers conduct studies on downstream tasks by using LLMs.

Sargent et al. [26] stated that large volumes of unclassified legal information are difficult to research and analyze. To tackle this challenge, they proposed

Table 1. Comparison of Research Approaches by Task and Model Type

Task	Model Type	Representative Studies
Classification	Non-LLM	[8,9,23]
	LLM	[15,26]
Information Extraction	Non-LLM	[7,17,20]
	LLM	[5,10]
Summarization	Non-LLM	[4]
	LLM	[3]
Information Retrieval	Non-LLM	[30,31]
	LLM	[14,35]
Legal Judgment Prediction	Non-LLM	[42]
	LLM	[25,36]

to establish a new legal taxonomy that accurately classifies summary judgment decisions of United Kingdom according to specific legal issues. They benefit from the potential of large language models for topic classification on the summary judgment decisions. Since the traditional models and methods are inadequate for capturing the complex and nuanced nature of legal language, the power of large language models to capture deep meanings is emphasized [15]. Izzidien et al. [15] highlighted the potential of large language models in classification by achieving 87.13% accuracy with formulation of appropriate prompts. Closed-set prompting strategies [38] are adopted with a predetermined topic list. Thereby, accuracy was increased by limiting the inference to a specific focal point. They also introduced advanced prompt approaches with few-shot learning strategies [2] by guiding the model with examples and constraints. They demonstrated the applicability of LLMs in legal domains through solely prompt formulation, without any fine-tuning [12].

After getting trained on high volume of text from various resources, LLM models can be useful to generate coherent and contextually relevant text based on given prompts [10]. Owing to the formulate prompts, LLMs can be used for different tasks without an additional training process. Ribeiro et al. [10] aimed to extract eight different legal elements (conclusion, subject matter, reasoning, etc.) from UK Employment Court decisions. Breton et al. [5] also compared large language models with classical BERT-based and rule-based methods for the information extraction tasks. They are demonstrated that more context-sensitive and flexible inferences can be made compared to traditional methods, especially since prompt strategies such as one-shot and zero-shot [2] are used. Both studies show that the context-capturing and flexible understanding capabilities of LLMs outperform traditional methods on the information extraction task. They proposed methods for the challenges in the legal domain by formulating prompt strategies to information extraction task without the need for a training process. In contrast, Baneddetto et al. [3] stated that while the LLMs are multilingual,

most of the training samples are in English. For instance, merely 0.11% of the training samples used for Llama 2 [32] are in Italian. They deals the challenge of summarizing in Italian languages while preserving the legal context although as most existing summarization systems are English-based. To overcome this challenge, they carried out a study with fine-tuning [12] to LLM architectures. In this studies, the fine-tuned models clearly outperform the results assessed by both ROUGE [22] and human evaluations. The results demonstrated that LLMs are powerful tools for summarizing information-dense texts in legal area.

The reasoning capability of LLMs has paved the way for development of new approaches to address numerous challenges in NLP. Unlike traditional information retrieval methods which rely solely on textual similarity, they have approached case selection as a semantic matching problem by enriching it with causal inferences [35]. In this study, first, the case texts were converted into semantic representations using LLM. Afterwards, the causal approach is discussed with a comparative decision mechanism and the articles of law affecting the outcome of the case with the help of the model. Therefore, by taking advantage of the deep understanding capabilities of the LLMs, they not only retrieved the given legal case descriptions but also produced interpretable descriptions for the retrieved documents. Thus, they closed the gap between retrieval performance and interpretability of retrieval causes. Huang et al. [14] also took a similar approach, by relying on the semantic understanding of LLMs to select related and unrelated information from related documents. In this study, continual training [11] and fine-tuning [12] methods were used to provide legal domain knowledge to general purpose trained models. These studies have revealed novel approaches for the information retrieval tasks by exploiting the reasoning capabilities of LLMs.

With inferential reasoning capabilities, LLMs have the potential to perform well on tasks such as legal decision prediction Legal Judgment Prediction (LJP) [25,36]. Wei et al. [36] used the CoT [37], which is an LLM technique with a step-by-step thought process, to extract critical elements such as important events, personals, and circumstances in an annotated manner. They used the extracted information for training of the Deep Neural Network (DNN) model they proposed. Consequently, the proposed model learns the final judgment decisions while also ensuring logical consistency. These inferences increase the interpretability of the model because it has demonstrated not only the model's decision predictions but also the reasoning behind the decisions. Peng et al. [25] used the retrieval augmentation generation (RAG) [21] method to improve the performance of the LJP and the consistency of the predictions. They have created a database to keep information about the indictments with the Athena framework they proposed, and integrated a semantic search mechanism into the model by using vectorization. Thereby, to tackle the lack of internal knowledge of LLMs, they have brought relevant legal information from the external knowledge base, and have included the information in the estimation process of the model. They demonstrated a significant performance increase in LJP tasks in the evaluations.

4 Data Resources in the Field of Law for Large Language Models

In law domain, large data resources can be accessed directly because legal documents and litigation processes are publicly available. The United Kingdom published the Cambridge Law Corpus (CLC) [44], which is a comprehensive open-sourced dataset, to support legal AI research. The dataset includes 250,000 UK court decisions and incorporates raw text, rich metadata and expert annotations. This dataset is a great resource for studies on NLP in legal texts [10,15,26]. By leveraging LLMs, CLC dataset was used as analysis and test dataset to carry out document analysis [15] and topics classification [26] tasks, respectively. Ribeiro et al. [10] preferred the $UKET_{ori}$ dataset, which is a subset of CLC, for their study. From the original $UKET_{ori}$ dataset which comprises of 52.339 cases, they have chosen a subset of 260 cases to use stratified sampling to achieve statistically robust findings.

LLMs demonstrate generalized inference performance due to training on large-scale datasets. Accordingly, several studies have demonstrated that downstream tasks could be addressed with prompting strategies without model training [2,15,26]. However, Benedetto et al. [3] asserted that models need to be fine-tuned in areas such as the legal field where the language is semantically difficult to understand and have complex language structures. Training LLMs has been known to require vast amounts of data [33]; however, field knowledge can be learned with a relatively smaller dataset by utilizing advanced fine-tuning methods [3]. They collected Italian Legal News to fine-tune the LLM. There are 3296 articles in total and the number of news articles significantly varies across different areas. They aimed to increase the model's knowledge of Italian law through fine-tuning methods. Furthermore, Breton et al. [5] used a small dataset where 1339 expressions were identified and annotated with French expressions from 200 documents selected by experts from the Luxembourg Traffic Code [29]. They used the dataset for testing and fine-tuning operations. The performance of fine-tuned models were evaluated on relatively small data sets by [3,5]. In both studies, an increase in the performance of models has been observed.

As a larger dataset with more than 2.6 million criminal cases, CAIL2018 [39], has been valuable a data resource for many studies [14,25,35]. Additionally, comprehensive evaluations has been performed by using subsets created from the CAIL2018 dataset [25,35]. Thus, it has been considered as a major resource for the Information Retrieval task in law domain. On the other hand, Huang et al. [14] used the Chinese Legal Corpus dataset, which was sampled from many datasets by the authors, to increase the legal knowledge of their model. In addition, they extended the JEC-QA [43] dataset with additional explanations to evaluate model performance.

Wang et al. [35] used the COLIEE [18] dataset as an evaluation resource in their study. The COLIEE is a publicly available dataset established by the COLIEE competition, which was organized at the International Conference on Artificial Intelligence and Law (ICAIL). They preferred using a subset of the COLIEE dataset to retrieve relevant articles from the Japanese Civil Code. The

test set included 319 new questions, where participants were given the task of retrieving relevant articles from 768 Civil Code articles. This dataset was used to compare LLMs with traditional models. In [36], a subset of the Chinese private lending cases dataset [24] was utilized for model training and testing. The dataset includes a total of 70.485 records. Reject, Partially Support, and Support have been used as the judgment labels for the claims in this dataset.

As a summary, it can be concluded that seven distinct data resources have been utilized by LLMs in the legal domain. Table 2 presents a list of these data resources and their usage purposes by LLMs. Cambridge Law Corpus and CAIL2018 have been observed as the most widely used data resources, as each of them is utilized by three different research studies. Besides, testing is determined as the most common purpose of usage with eight studies, followed by training with five studies.

Table 2. The Data Resources of LLMs in Legal Domain and Their Usage Purposes

Reference	Data Resources	Purpose of Usage
[10]	Cambridge Law Corpus	Testing
[15]	Cambridge Law Corpus	Document Analysis
[26]	Cambridge Law Corpus	Testing
[3]	Custom Data	Training, Testing
[5]	Custom Data	Training, Testing
[14]	CAIL2018, JEC-QA	Training, Testing
[25]	CAIL2018	Testing
[35]	CAIL2018, COLIEE	Testing
[36]	[24]	Training, Testing

5 Evaluation Strategies for LLMs in Law Domain

As more complex AI methods emerge, evaluating task performance becomes more difficult. Hence, evaluating LLMs is quite a challenge on its own. Several studies especially focused on addressing the evaluation of the tasks in legal domain. Traditional performance evaluation metrics such as Accuracy, Precision, Recall and F1-score are widely used [15,25,26,36]. Classification tasks are evaluated in a direct comparative manner on labeled data. LJP tasks are evaluated in a manner similar to classification tasks, which are assessed directly using labeled data [5,36]. Wang et al. [36] also addressed the judgment selection as a binary classification problem. They preferred classification accuracy as a criterion to evaluate their comparison algorithms. Breton et al. [5] presented a novel approach to address the inherent limitations of traditional evaluation metrics when

applied to LLMs for the information retrieval tasks. To overcome these limitations, an advanced evaluation framework that incorporates a detailed analysis of True Positive examples and the combined use of the Levenshtein distance [40] was presented.

Generator outputs for the summarization task can not be evaluated directly as a match between the ground truth and the prediction. Instead, quantitative metrics such as ROUGE [22] and BERTScore [41] are preferred [3]. ROUGE measure the overlap between automatic summaries and reference summaries to assess their quality. On the other hand, BERTScore evaluates the semantic similarity of texts and measures the degree to which the summaries produced by the model overlap with the reference summaries.

Although there are many evaluation criteria, manual evaluation with experts is still widely used [3,10,14,26]. These studies interpret model performance based on the evaluations of field experts. Furthermore, the hallucination problem of LLMs is a major problem widely stated by researchers. The hallucination problem in LLMs refers to the generation of plausible-sounding but factually incorrect or entirely fabricated content. The assessment hallucinations is a major assessment challenge addressed by current studies, and measuring hallucinations in the predictions is utilized as an assessment criterion [13,14]. Hu et al. [13] examined hallucination evaluation for LLMs with their benchmark called LegalHalBench. They proposed the Non-Hallucination Status Rate method which defines a comparative rate for hallucination in LLMs.

6 Discussion

Addressing the tasks with natural language processing techniques in the field of law is quite challenging due to the complex and nuanced nature of legal language [15]. The studies using LLMs capture deep meanings to overcome the challenging tasks listed in Table 1. However, relying solely on the generalized training of LLMs proves inadequate to address complex tasks [3,14]. Moreover, since the majority of training data for LLMs is composed of English texts, their capacity to interpret legal documents in other languages remains limited [3]. Therefore, training models to increase knowledge in the legal domain appears to enhance task performance [3].

Table 3 provides an overview of studies utilizing Large Language Models in legal domain. It is observed that the Cambridge Law Corpus and the CAIL2018 datasets are the most commonly employed with three studies each. Cambridge Law Corpus and the CAIL2018 are utilized in [10,14,15,25,26,35] respectively. Furthermore, in terms of evaluation criteria, human evaluation appears to be a frequently adopted approach [3,10,14,26], with researchers often relying on domain experts to assess model outputs. The review of the studies reveals that evaluating Large Language Models in the legal domain remains a particular challenge.

It has been demonstrated by [35,36] that the reasoning capabilities of LLMs have great potential to overcome challenges in the legal domain. While LLMs are

Table 3. Overview of Studies in Legal Domain Using Large Language Models

Article	Task	Data Resource(s)	Evaluation Criteria
[10]	Information Extraction	Cambridge Law Corpus	Human Evaluation
[5]	Information Extraction	Custom Data	F1 Score, Recall, Precision
[15]	Classification	Cambridge Law Corpus	F1 Score
[26]	Classification	Cambridge Law Corpus	F1 Score, Human Evaluation
[3]	Summarization	Custom Data	ROUGE, BERTScore, Human Evaluation
[14]	Information Retrieval	CAIL2018, JEC-QA	Accuracy, Human Evaluation, Non-Hallucination Status Rate
[35]	Information Retrieval	CAIL2018, COLIEE	Accuracy
[25]	Legal Judgment Prediction	CAIL2018	Accuracy
[36]	Legal Judgment Prediction	[24]	F1 Score, Recall, Precision

rarely adopted as base models in research, they commonly function as auxiliary components or decision-making aids. However, LLMs cause concerns in tasks due to hallucination problems and lack of interpretability. In domains such as law, where the reasoning behind justification is of critical importance, the interpretability of artificial intelligence model outputs is crucial. Novel approaches to improve the interpretability of models have been proposed to address the lack of interpretability [36,37]. Besides, approaches that benefit from external resources such as RAG have been observed to prevent hallucinations and increase the consistency of the models [25].

Especially in decision-making positions that directly affect human life, such as in the field of law, LLMs pose a significant risk to direct people to false information. Addressing concerns of hallucination, bias, and lack of interpretability in LLMs is a major requirement for AI technologies in law. In addition, the basic rules of the law which vary depending on the region, country and situation make it difficult to propose a general evaluation method for LLMs in the field of law.

7 Conclusion

In this paper, we discussed the tasks addressed with natural language processing techniques in the legal domain and examined the novel approaches developed with LLMs for these tasks. The deep contextual understanding capabilities of LLMs have been demonstrated to have great potential in tasks which require high reasoning and capture semantically challenging language structures in the legal domain. Novel approaches for legal domain emerge from particularly the reasoning capabilities of LLMs. Additionally, it has been demonstrated that LLMs are more successful in performance than traditional methods. Although LLMs can be applied to legal challenges without requiring training, fine-tuned models have been established to be more successful in grasping the deep context of legal texts.

Hallucination, bias, and lack of interpretability in the LLMs may raise serious concerns. Reducing hallucination and addressing ethical concerns are of crucial

importance to increase the application of LLMs in the field of law. Furthermore, the lack of a standard for evaluation and the difficulty of assessing the hallucinations constitute major challenges for application of LLMs in the legal domain. Mitigating these concerns requires the establishment of robust, high-quality datasets and standardized benchmarks. Therefore, generating enhanced datasets and developing new benchmarks are essential steps toward improving the trustworthiness and applicability of LLMs in legal contexts.

References

1. Abanoz, B.: Hukuki bilgiye dijital erişim. Marmara Üniversitesi Hukuk Fakültesi Hukuk Araştırmaları Dergisi **26**(1), 190–203 (2020)
2. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., Sontag, D.: Large language models are few-shot clinical information extractors (2022). https://arxiv.org/abs/2205.12689
3. Benedetto, I., Cagliero, L., Ferro, M., Tarasconi, F., Bernini, C., Giacalone, G.: Leveraging large language models for abstractive summarization of italian legal news. Artif. Intell. Law 1–21 (2025)
4. Benedetto, I., La Quatra, M., Cagliero, L.: LegITBART: a summarization model for Italian legal documents. Artif. Intell. Law 1–31 (2025)
5. Breton, J., et al.: Leveraging LLMs for legal terms extraction with limited annotated data. Artif. Intell. Law 1–27 (2025)
6. Brown, T.B., et al.: Language models are few-shot learners (2020). https://arxiv.org/abs/2005.14165
7. Çetindağ, C., Yazıcıoğlu, B., Koç, A.: Named-entity recognition in Turkish legal texts. Nat. Lang. Eng. **29**(3), 615–642 (2023)
8. Chen, H., Wu, L., Chen, J., Lu, W., Ding, J.: A comparative study of automated legal text classification using random forests and deep learning. Inform. Process. Manage. **59**(2), 102798 (2022)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018). http://arxiv.org/abs/1810.04805
10. Ribeiro de Faria, J., Xie, H., Steffek, F.: Information extraction from employment tribunal judgments using a large language model. Artif. Intell. Law 1–22 (2025)
11. Gururangan, S., et al.: Don't stop pretraining: adapt language models to domains and tasks (2020). https://arxiv.org/abs/2004.10964
12. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification (2018). https://arxiv.org/abs/1801.06146
13. Hu, Y., Gan, L., Xiao, W., Kuang, K., Wu, F.: Fine-tuning large language models for improving factuality in legal question answering. arXiv preprint arXiv:2501.06521 (2025)
14. Huang, Q., et al.: Lawyer llama technical report. arXiv preprint arXiv:2305.15062 (2023)
15. Izzidien, A., Sargeant, H., Steffek, F.: LLM vs. lawyers: identifying a subset of summary judgments in a large UK case law dataset. arXiv preprint arXiv:2403.04791 (2024)
16. Jacobs, G.L.: Legal technology: present and future trends. Legal Econ. **13**, 29 (1987)

17. Ji, D., Tao, P., Fei, H., Ren, Y.: An end-to-end joint model for evidence information extraction from court record document. Inform. Process. Manage. **57**(6), 102305 (2020)
18. Kim, M.Y., Rabelo, J., Goebel, R., Yoshioka, M., Kano, Y., Satoh, K.: Coliee 2022 summary: Methods for legal document retrieval and entailment. In: JSAI International Symposium on Artificial Intelligence, pp. 51–67. Springer (2022)
19. Küçük, D., Can, F.: Hukuki metinlerin otomatik işlenmesinde yapay zekâ teknolojilerinin kullanimi. Bilişim Hukuku Dergisi **6**(1), 1–23 (2024)
20. Leitner, E., Rehm, G., Moreno-Schneider, J.: Fine-grained named entity recognition in legal documents. In: International Conference on Semantic Systems, pp. 272–287. Springer (2019)
21. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks (2021). https://arxiv.org/abs/2005.11401
22. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out: Proceedings of the ACL-04 workshop, pp. 74–81 (2004)
23. Mumcuoğlu, E., Öztürk, C.E., Ozaktas, H.M., Koç, A.: Natural language processing in law: prediction of outcomes in the higher courts of turkey. Inform. Process. Manage. **58**(5), 102684 (2021)
24. MLY NLP: LJP-MSJudge: a multi-scale dataset for legal judgment prediction. https://github.com/mly-nlp/LJP-MSJudge (2023). Accessed 26 March 2025
25. Peng, X., Chen, L.: Athena: Retrieval-augmented legal judgment prediction with large language models (2024). https://arxiv.org/abs/2410.11195
26. Sargeant, H., Izzidien, A., Steffek, F.: Topic classification of case law using a large language model and a new taxonomy for UK law: AI insights into summary judgment. Artificial Intelligence and Law, pp. 1–49 (2025)
27. Shaheen, Z., Wohlgenannt, G., Filtz, E.: Large scale legal text classification using transformer models. arXiv preprint arXiv:2010.12871 (2020)
28. Shukla, A., et al.: Legal case document summarization: Extractive and abstractive methods and their evaluation (2022). https://arxiv.org/abs/2210.07544
29. Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Dann, J.: Automated extraction of semantic legal metadata using natural language processing. In: 2018 IEEE 26th International Requirements Engineering Conference (RE), pp. 124–135 (2018). https://doi.org/10.1109/RE.2018.00022
30. Tagarelli, A., Simeri, A.: Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code. Artif. Intell. Law **30**(3), 417–473 (2022)
31. Thomas, M., et al.: Quick check: a legal research recommendation system. In: NLLP@ KDD, pp. 57–60 (2020)
32. Touvron, H., et al.: Llama 2: Open foundation and fine-tuned chat models (2023). https://arxiv.org/abs/2307.09288
33. Trozze, A., Davies, T., Kleinberg, B.: Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers? Artif. Intell. Law 1–47 (2024)
34. Vatsal, S., Meyers, A., Ortega, J.E.: Classification of us supreme court cases using BERT-based techniques. arXiv preprint arXiv:2304.08649 (2023)
35. Wang, Z., Ding, Y., Wu, C., Guo, Y., Zhou, W.: Causality-inspired legal provision selection with large language model-based explanation. Artif. Intell. Law 1–25 (2024)
36. Wei, B., Yu, Y., Gan, L., Wu, F.: An LLMs-based neuro-symbolic legal judgment prediction framework for civil cases. Artif. Intell. Law 1–35 (2025)

37. Wei, J., et al.: Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 (2022)
38. White, J., et al.: A prompt pattern catalog to enhance prompt engineering with Chatgpt (2023). https://arxiv.org/abs/2302.11382
39. Xiao, C., et al.: Cail2018: A large-scale legal dataset for judgment prediction (2018). https://arxiv.org/abs/1807.02478
40. Yujian, L., Bo, L.: A normalized levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1091–1095 (2007)
41. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
42. Zhong, H., Wang, Y., Tu, C., Zhang, T., Liu, Z., Sun, M.: Iteratively questioning and answering for interpretable legal judgment prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 1250–1257 (2020)
43. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: JEC-GA: a legal-domain question answering dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9701–9708 (2020)
44. Östling, A., et al.: The Cambridge law corpus: A dataset for legal AI research (2024). https://arxiv.org/abs/2309.12269