Large language models for automated Q&A involving legal documents: a survey on algorithms, frameworks and applications

Xiaoxian Yang, Zhifeng Wang, Qi Wang, Ke Wei and Kaiqi Zhang
Shanghai Polytechnic University, Shanghai, China, and

Jiangang Shi
Shanghai Shang Da Hai Run Information System Co., Ltd, Shanghai, China

Received 18 December 2023
Revised 26 January 2024
16 February 2024
Accepted 18 February 2024

Abstract

Purpose – This study aims to adopt a systematic review approach to examine the existing literature on law and LLMs.It involves analyzing and synthesizing relevant research papers, reports and scholarly articles that discuss the use of LLMs in the legal domain. The review encompasses various aspects, including an analysis of LLMs, legal natural language processing (NLP), model tuning techniques, data processing strategies and frameworks for addressing the challenges associated with legal question-and-answer (Q&A) systems. Additionally, the study explores potential applications and services that can benefit from the integration of LLMs in the field of intelligent justice.

Design/methodology/approach – This paper surveys the state-of-the-art research on law LLMs and their application in the field of intelligent justice. The study aims to identify the challenges associated with developing Q&A systems based on LLMs and explores potential directions for future research and development. The ultimate goal is to contribute to the advancement of intelligent justice by effectively leveraging LLMs.

Findings – To effectively apply a law LLM, systematic research on LLM, legal NLP and model adjustment technology is required.

Originality/value – This study contributes to the field of intelligent justice by providing a comprehensive review of the current state of research on law LLMs.

Keywords Large language model, Legal documents, Natural language processing, Automated Q&A application

Paper type Research paper

1. Introduction

In recent years, an increasing number of countries have made policy decisions to incorporate artificial intelligence (AI) into broader judicial reform initiatives. The impetus behind judicial reform and the growing interest in leveraging AI are motivated by the chronic challenges facing courts (Wang, 2021), such as the exponential increase in the amount of required casework and a shortage of qualified professionals within the judiciary system. To solve these challenges, the construction of intelligent justice methods (Wang and Tian, 2023) has proven to be an effective approach. To realize the objective of intelligent justice, swift, cost-free and precise legal consultation services must be offered. Additionally, enhancing the caliber of the provided legal services, reducing user consultations and lawyer service expenses and addressing the disparity between the supply and demand for legal advisory services are

International Journal of Web Information Systems

1744-0084
DOI 10.1108/IJWIS-12-2023-0256

essential pursuits. Consequently, substantial practical significance and application value can be derived from the study of automatic legal question-and-answer (Q&A) methods.

Although automatic legal Q&A methods have undergone extensive investigation, no satisfactory legal Q&A systems based on large language models (LLMs) (Taulli, 2023) are available. LLMs belong to an important branch of natural language processing (NLP). In particular, they are emerging as a promising paradigm that is capable of processing vast amounts of text data and extracting rich semantic and grammatical knowledge from them. By pretraining a model on a large-scale text corpus, the inherent laws of language can be learned and transferred to downstream NLP tasks.

The applications of law LLMs include legal document research, automatic legal Q&A tasks and legal document generation. LLMs possess formidable language processing capabilities, enabling lawyers and researchers to swiftly gather and analyze numerous legal documents and cases. They can efficiently provide accurate information and background knowledge. Moreover, LLMs can provide advice and solutions to lawyers by delivering personalized insights derived from the understanding of case details and relevant regulations. Currently, one of the foremost concerns for researchers is the efficient utilization of resources to construct a high-performance LLM. Specifically, the existing research focuses on the following three aspects: fine-tuning model algorithms, optimizing training data and refining overall frameworks.

In this paper, we conduct a thorough law LLM survey. From data to applications, we provide a bottom-up systematic review of the existing work, particularly focusing on model fine-tuning. First, this paper introduces the basic concepts of LLMs and reviews the related research on legal NLP. Second, this paper provides an in-depth summary of the research conducted on fine-tuning law LLMs considering their algorithms, frameworks and data. Third, the applications, challenges and key issues concerning law LLMs are discussed. A systematic outline of this survey is presented in Figure 1.

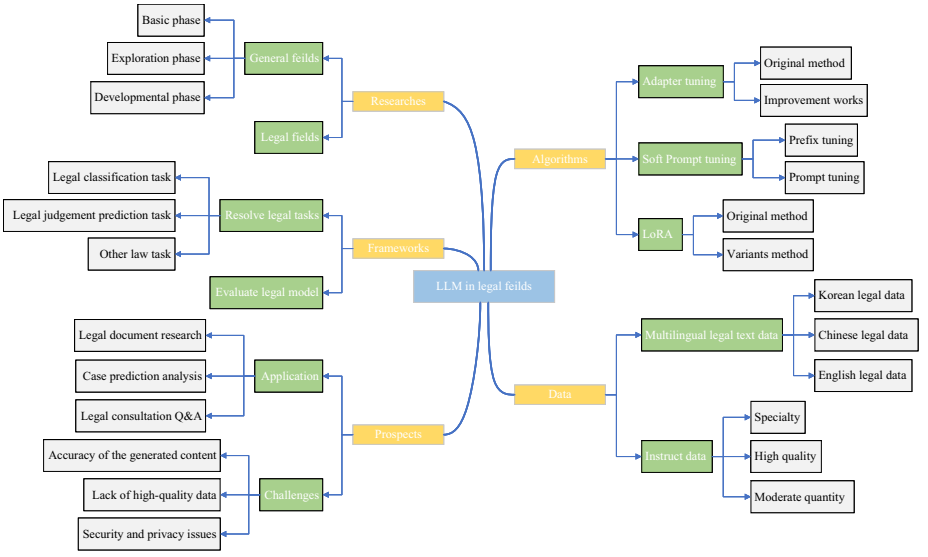


Figure 1.
Complete overview of this paper

Source: Author's creation

In conclusion, this article provides an overview and summary of the research conducted on the comprehensive adaptation and applications of large language models in the legal domain. While notable progress has been made regarding the intersection of large language models and legal artificial intelligence, as depicted in Table 1, limited attention has been given to adapting large language models specifically tailored to the legal domain. Hence, this paper delves into the recent advancements achieved in terms of leveraging large language models within the legal domain and investigates the applications of these models from four key perspectives: model adaptation algorithms, data processing strategies, frameworks and application services.

The remainder of this paper is organized as follows. In Section 2, we review the concepts and development trends related to LLMs. Then, we survey the existing work on fine-tuned LLMs in Section 3. Case studies concerning the applications of law LLMs are selected and discussed in Section 4. Finally, we summarize the paper in Section 5.

2. Large language models and legal natural language processing

2.1 The requirements for combining legal documents and large language models

2.1.1 The current judicial dilemma. With the construction of a society based on the rule of law and the enhancement of citizens' legal awareness levels, the size of the population in need of legal services continues to rise. However, due to the improper allocation of lawyer resources and the small growth rates exhibited by the scale of lawyers and the number of available personnel, users may face difficulties when attempting to obtain legal representation and long waiting times when encountering legal issues (Qu *et al.*, 2023). Specifically, legal advisory services may pose some of the following challenges and limitations.

- Many legal documents cover various aspects of life. If these documents are only manually interpreted and applied, a significant amount of human resources is needed to provide explanations and guidance to counselors.
- Human memory is prone to bias, which may result in omissions or misinterpretations of contents when interpreting and applying legal documents. The ultimate outcome may be a mismatch between the applicable legal documents and the actual situation, leading to legal disputes and controversies.
- The language and structures of legal documents are usually more formal and professional than those of other documents (Gordon and Breaux, 2014). Ordinary users usually lack professional knowledge in the legal field and may not be able to fully understand the meaning and scope of legal documents.

To overcome these challenges, the application of AI technology has become a potential solution. By using NLP techniques, AI can automatically parse and understand legal

Works	Focus
Lai *et al.* (2023)	Law LLMs, challenges, applications
Zhong *et al.* (2020a)	Legal artificial intelligence, applications
Dias *et al.* (2022)	Artificial intelligence revolutionizing the legal field
Sheik and Nirmala (2021)	Deep learning models sequence-to-sequence neural network architectures
Our Paper	Model tuning algorithms, data processing strategies frameworks, application services

Source: Authors' creation

Table 1. Existing survey work on legal AI and LLMs

documents, providing accurate guidance to consultants (Ma, 2022). From a deeper perspective, AI can be used to conduct high-quality legal document searches to optimize and improve the case-handling process and even assist in making judicial judgments.

2.1.2 The advantages of large language models in judicial scenarios. LLMs represent a novel and crucial research technology within the AI domain. By training on a comprehensive text corpus, an LLM learns the inner patterns of language and acquires a massive amount of linguistic knowledge and strong contextual understanding capabilities. LLMs have the ability to analyze questions presented in natural language, and they can also generate relevant and accurate answers. Compared to traditional rule-based or keyword-matching-based systems (Katz *et al.*, 2023), LLMs can understand more complex semantics and contexts to better solve questions posed by users. In the legal field, LLMs can cite relevant legal documents, cases, legal academic research and regulations. This capability of LLMs can help users better understand legal documents, their interpretations and the relevant contexts. This approach provides more authoritative and reliable answers for users. In addition, LLMs can learn from users' feedback and behavioral patterns to continually optimize the accuracy of their answers. By analyzing users' questioning styles, concerns and feedback, models can adapt and provide answers that are more responsive to the users' needs.

2.2 The concepts and development trends of large language models

At the end of 2022, OpenAI released ChatGPT, an LLM application that attracted widespread attention due to its excellent ability to respond to human requests described in natural language. ChatGPT was built based on a generative pre-trained transformer (GPT) model, and it incorporates deep learning, instruction tuning, multi-task learning and other technologies. Case studies are among the most direct and effective ways to test whether new technologies are valuable, and this is also true for LLMs. According to their historical release times, we present several of the existing LLMs. Figure 2 shows the representative LLMs released during each period.

2.2.1 Basic phase. The research conducted in the basic stage focused mainly on model architectures. Vaswani *et al.* (2017) proposed a new type of neural network model structure called a transformer. Almost all LLMs since then have been developed based on this architecture. In 2018, OpenAI released the first GPT model called GPT-1 (Radford *et al.*, 2018). It adopts a hybrid method with unsupervised training and supervised fine-tuning. Subsequently, GPT-2 (Hauser *et al.*, 2002) was released in 2019. The researchers introduced the concept of multitask learning, and a novel method was proposed to execute tasks by performing unsupervised language modeling without implementing an explicit fine-tuning process using labeled data.

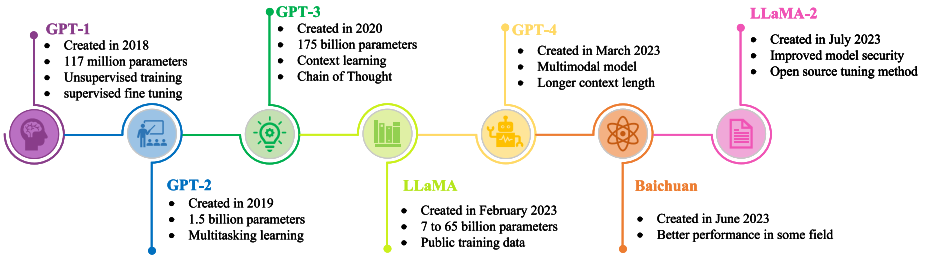


Figure 2.
The history of LLMs

Source: Author's creation

2.2.2 Exploration phase. The exploration phase was dedicated to improving the output capabilities of models. To further improve the performance of a model in cases with few or zero samples, GPT-3 (Brown *et al.*, 2020) was released. Wei *et al.* (2021) proposed the instruction tuning method, which unifies many tasks into a generative natural language understanding framework. Brown *et al.* (2020) proposed the in-context learning method. This approach splices a small number of labeled instances for different tasks into samples. Wei *et al.* (2022) proposed the chain-of-thought prompting strategy, which enables the constructed model to use a prompting mechanism containing intermediate reasoning steps to solve a task.

2.2.3 Developmental phase. The development stage of LLMs started with the release of the GPT-4 (OpenAI, 2023) model. In 2023, OpenAI released this tool and expanded its accepted inputs from text to multiple data types. As the research on LLMs continued to deepen, Meta released LLaMA (Touvron *et al.*, 2023a) in February 2023. LLaMA is an open source and efficient LLM with between 7 billion and 65 billion parameters. In June, Yang *et al.* (2023) released Baichuan, which is a Chinese and English LLM with between 7 billion and 13 billion parameters. Then, in July 2023, Meta released LLaMA-2 (Touvron *et al.*, 2023b). The dialog-related use cases of the model were optimized, with a greater emphasis on safety and usefulness than in the previous version of the model.

In recent years, LLMs have been rapidly developed and proliferated in the general domain. While these models have demonstrated remarkable capabilities across various fields, ranging from text generation to semantic understanding, their performance in terms of addressing legal issues has not met expectations. The complexity of the legal domain cannot be underestimated, as it involves the interpretation of legal texts, case analyses, legal reasoning scenarios and more, demanding precise language, rigorous logic and a deep understanding of legal contexts. LLMs have deficiencies in this regard for several reasons. First, legal documents and knowledge systems are relatively closed and specialized, and the training data used for general LLMs often lack comprehensive coverage of sufficient legal terminology and concepts, resulting in biases and inaccuracies with respect to understanding legal terms and expressing legal concepts. Second, legal issues often require nuanced logical reasoning and meticulous analyses of facts, which are areas in which general LLMs have demonstrated weaker performance, leading to insufficient or erroneous reasoning. Furthermore, legal documents require the consideration of various possible interpretations and scenarios; this poses challenges to general language models in terms of effectively handling such complexities. To address these challenges, fine-tuning general language models is a viable approach.

2.3 Research on legal natural language processing

2.3.1 Model studies. To test the effectiveness of pre-trained language models in the legal domain, Song *et al.* (2022) conducted a thorough empirical evaluation of the pre-trained models developed for legal NLP. They studied approximately 8 representative legal data sets, with sample sizes ranging from 900 to 57 K. These data sets covered a spectrum consisting of five NLP tasks, encompassing binary classification, multilabel classification, multiple-choice Q&A, summarization and information retrieval. Nguyen (2023) designed a virtual legal assistant built on a version of the GPT-3 language model that was fine-tuned for the legal domain. The system was designed to provide legal assistance to users in a conversational manner, helping them with tasks such as answering legal questions, generating legal documents and providing legal advice. Huang *et al.* (2023a) proposed a new framework to adapt LLMs to specific domains and built Lawyer LLaMA, an LLM for the legal domain, based on this framework. First, legal domain knowledge was continuously

injected during the pretraining stage, and the model was taught to learn legal professional skills using properly designed and supervised fine-tuning tasks. Second, to alleviate the hallucinations that may occur when the model generates content, a retrieval module was added to extract relevant legal articles before the model answers any query. Lee (2023) designed a language model specifically for the legal domain. Their study presented the development trends of LexGPT models based on GPT-J models and pre-trained with the Pile of Law. First, their model was further trained using human feedback-based learning methods. Second, professional data fine-tuning models were used, so the code of the model itself did not need to be modified during this process. Thus, legal professionals can create a legal language model for downstream tasks with minimal effort and professional technical knowledge.

2.3.2 Research on unique legal issues. A substantial portion of the online data available for the legal domain consists of unstructured textual information. These data heavily rely on information stored in various legal documents. The legal domain is renowned for its complexity, and its processes often rely on human experts for knowledge interpretation. Establishing the relevance and similarity between two cases based on the expositions found in various legal documents is a common yet nontrivial task performed by legal experts. Wagh and Anand (2017) explored the applicability of network analysis to compare two approaches for identifying the similarities among legal documents. To address the issues regarding the excessive pending cases and conflicting rulings within the Brazilian Judiciary, judges on the Brazilian Supreme Court often need to cite multiple existing precedents in their judgments. Resck *et al.* (2022) proposed a network-based visual analysis system, LegalVis, which was designed to support the analysis of legal documents that cite or may cite precedents. First, the problem of identifying latent citations, known as nonexplicit citations, was modeled as a classification problem. Second, the authors explored the relationships between documents and their precedents. LegalVis was designed with three interactive visual components. The first component displays an overview of the input data over time. The second component allows the constructed system to filter and group related documents by topic. The last module is a component used to display a text document designed to interpret the model output by indicating which passages may refer to precedents. The system was designed to enable judicial experts to quickly identify and analyze these precedents. Given the impact of the law on society, judges must review cases annually to ensure the accurate application of legal documents. However, case reviews also raise questions regarding privacy when accessing documents from other courts. The safeguarding of personal private data remains challenging. El Moussaoui *et al.* (2023) developed a system consisting of two components to generate legal judgments. The first component is a data extraction model that recognizes key entities. The second component contains a Python module that hides the personal information extracted by the first component. Together, these components form an anonymous generative judgment system.

3. Fine-tuning large language models for the legal domain

3.1 Algorithms

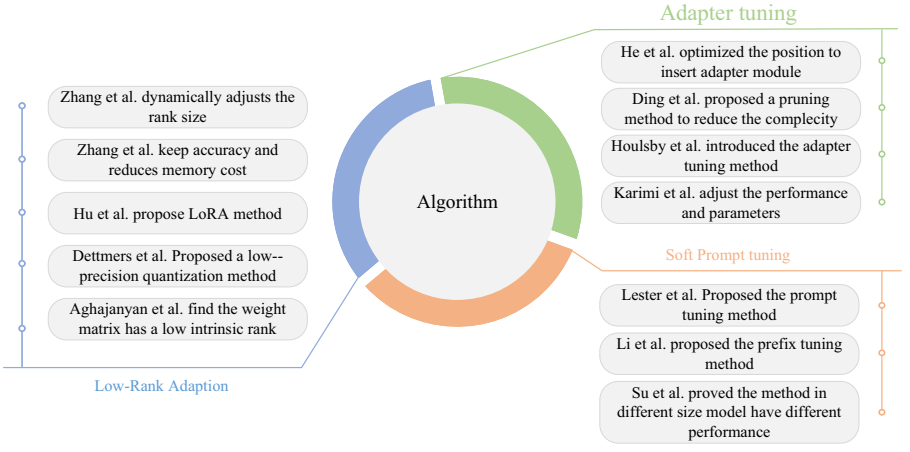
After the pretraining phase, LLMs demonstrate relatively remarkable performance across various NLP tasks. However, fine-tuning a pre-trained model on a downstream data set has been shown to yield superior performance. Owing to the large number of parameters contained in an LLM, fine-tuning all of its parameters incurs a substantial computational cost when the model is applied to downstream tasks. For this reason, the existing algorithms mainly focus on efficiently fine-tuning parameters. The goal is to successfully adapt a model by training a small number of parameters. The research community collectively refers to

such methods as parameter-efficient fine-tuning (PEFT) (Ding *et al.*, 2023; Sabry and Belz, 2023) approaches. PEFT fixes or freezes most of the parameters of a pretraining model and fine-tunes only a small number of additional parameters. It can downscale the training parameters and the memory footprint of the used GPU. In this section, we provide a detailed overview of three PEFT methods: adapter tuning, soft prompt tuning and low-rank adaptation (LoRA). A systematic view of these algorithms is presented in Figure 3.

3.1.1 Adapter tuning. Adapter tuning is a crucial PEFT method that is used in LLM tuning research. Integrating an adapter module when processing downstream tasks enables the retention of most of the original model parameters throughout the fine-tuning process. Specifically, an adapter module consists of two feedforward networks (FFNs), a nonlinear layer and a residual connection layer. It compresses the original feature vector to a smaller dimensionality, then performs a nonlinear transformation and restores the vector to its original dimensionality. Concerning the insertion location, an adapter module is typically inserted into the model structure after its attention layer and feedforward layer.

Houlsby *et al.* (2019) introduced the adapter-tuning method, which implements model fine-tuning by adding a small neural network (an adapter module) to the used model. Compared with the traditional full-parameter fine-tuning method, the adapter-tuning method requires only a small number of parameters before the fine-tuning process. For example, on the GLUE benchmark data set (Wang *et al.*, 2018), the adapter-tuning method needs to train only approximately 3.6% of the parameters required by the bidirectional encoder representations from transformers (BERT) model. He *et al.* (2021) suggested that the insertion of adapter modules at different locations in a network architecture influences the performance of the constructed model. They optimized the adapter module insertion position, leading to improved results.

While the adapter-tuning method yields significantly enhanced computational efficiency, it is worth noting that the increased number of parameters in the adapter module may impact the overall accuracy of the associated model. To address this challenge, He *et al.* (2022) proposed a pruning method. This approach reduces the complexity of a model by identifying and eliminating the parameters that contribute less to its overall performance.



Source: Author's creation

Figure 3. State-of-the-art algorithms for LLM tuning

Consequently, this method enhances the inference speed of the model and effectively mitigates the issue of model overfitting. Karimi *et al.* (2021) used low-rank optimization and parameterized the hypercomplex multiplication process to adjust the balance between model performance and the number of trainable parameters. Low-rank decomposition was applied to the original parameter matrix, and parameterized hypercomplex multiplication was used for computational purposes. This approach enables researchers to operate in resource-constrained environments, facilitating more efficient model deployment and inference procedures.

The adapter-tuning method effectively reduces the number of training parameters required during the fine-tuning phase. This technique reduces the extensive computational resources and time needed to fine-tune an entire model, and it learns sufficient knowledge during the training process. The adapter-tuning method has been widely used in fine-tuning research involving LLMs for specific tasks.

3.1.2 Soft prompt tuning. Soft prompt tuning is a prevalent PEFT method. It adapts to downstream tasks by incorporating a unique, continuous and trainable prefix sequence into the input or intermediate state preceding a model. This method can be further divided into prompt tuning and prefix tuning approaches. The prompt tuning method splices a learnable tensor for a specific task in front of the embedding vector input by the used model. On the other hand, the prefix tuning method splices a continuous sequence for a specific task in front of the model embedding vector and the hidden state so that the model can quickly understand task-related information when processing the input.

When handling tasks, LLMs typically use sentences or fragments as inputs, generating outputs based on these inputs. However, this approach presents challenges, including potential deviations from the expected values in the generated output and a lack of precise control over the used content. Lester *et al.* (2021) proposed the prompt tuning method, which does not update the parameters of the underlying model. In addition, they mainly focused on setting prompt words or templates that could guide the model to generate the required output. The resulting model could more quickly understand and align task-related contextual information. Su *et al.* (2022) conducted an experiment on the performance of the prompt tuning method under prompt words with different lengths and models of different sizes. The results showed that the efficiency of using the prompt tuning method to fine-tune a model increased as the model size increased. Li and Liang (2021) proposed the prefix tuning method. The concepts of the prefix tuning and prompt tuning methods are closely related. During the experiment, the authors observed that directly optimizing the prefix sequence P could easily result in an unstable model training process. Therefore, an FFN was initially used to transform P, which was subsequently spliced with the embedding vector and hidden state. The prefix sequence and the learnable parameters in the FFN were optimized by using the gradient descent method during the training process. The results showed that in some NLP tasks, this method needs to fine-tune only 0.1% of the total number of parameters to achieve full fine-tuning effect.

Both the prompt tuning and prefix tuning methods are soft prompt tuning methods. These techniques add specific prompt words in front of the input sequence, enabling the used model to swiftly understand and align task-related contextual information. This approach is highly flexible and can be adapted to fine-tune models for various tasks. However, the maximum acceptable model length is limited; as the number of prompt parameters increases, the maximum length of the actual input sequence also decreases. This means that the model performance achieved by the soft prompt tuning method does not increase monotonically with the number of trainable parameters.

3.1.3 Low-rank adaptation. LoRA is a PEFT method that needs to train only a small number of parameters to produce good results when adapting LLMs to downstream tasks. This method adds a side branch network to the fully connected layer of the used model. It uses the product of two rank decomposition matrices to approximate the updated value of the fully connected layer weight during the knowledge domain adaptation process. Only the parameters in the side branch network are updated, while the weights of the pre-trained layer are kept constant.

To develop more PEFT methods, researchers have investigated model parameter-related solutions and conducted in-depth research on the parameter weight matrix. Aghajanyan *et al.* (2021) reported that after models were fine-tuned for specific tasks, the weight matrix usually had a low intrinsic rank. Therefore, Hu *et al.* (2021) proposed a LoRA method based on model weight matrix adjustment. First, they believed that even if the number of parameter updates were projected into a smaller subspace, the effectiveness of the learning process would not change. Second, this method retains the parameters contained in the pretraining model and uses only a low-rank matrix with few parameters to approximate the parameter update process of the model weight matrix. Third, only the low-rank matrix parameters are optimized during training. The main advantage of the LoRA method is that it greatly saves computer memory and storage costs. Aghajanyan *et al.* (2021) noted that by fine-tuning the GPT-3 model, when r equals four and bypasses only the two matrices of the attention module, the saved checkpoint memory changes from 350 GB to 35 MB. The GPU memory used during training is reduced from 1.2 TB to 350 GB, and the training speed increases by 25% compared to that of full-parameter fine-tuning. To improve the fine-tuning accuracies achieved under different tasks, Zhang *et al.* (2023) proposed the LoRA-FA method, which keeps the changing weight of the model in the low-rank space during the LLM fine-tuning process and removes the storage for the full-rank input activation step.

To further reduce the imposed computational burden, Dettmers *et al.* (2023) proposed the QLoRA method, which is a low-precision quantization and fine-tuning technology for deep neural networks. The addition of 4-bit normal float quantization and double quantization processes enables high-fidelity 4-bit fine-tuning. Paged optimizers were introduced to address the memory fluctuations caused by gradient checkpoints, preventing insufficient memory issues. This strategy does not modify the logic of the LoRA method but rather saves further computational overhead by quantizing the pretrained model to 4 bits. Zhang *et al.* (2022) proposed the AdaLoRA method, which dynamically adjusts the rank size according to the importance of each weight matrix to downstream tasks during the fine-tuning process. This method can further reduce the number of trainable parameters and yield improved model performance.

Table 2 presents a comprehensive comparative analysis of various model-tuning methodologies.

3.1.4 Fine-tuning case study. Fine-tuning an LLM is critical for aligning the model with a specific use case to improve the obtained results. Prominent pretrained LLMs can be fine-tuned and applied to specific tasks, especially in the legal domain. Fernandes *et al.* (2023) approached the challenge of categorizing legal foundations in first-degree sentences under Brazilian law by refining the multilingual MultiFit model through a unique foundational data set. The aim was to comprehensively train the model to precisely classify legal foundations. Wei *et al.* (2023) conducted a series of experiments that compared a standard, pretrained DistilBERT model with a fine-tuned DistilBERT model in the context of a downstream NLP task focused on text classification. By using domain-specific data extracted from actual legal cases, the study suggested that fine-tuning enhances the performance of LLM-based text classifiers. Malic *et al.* (2023) defined legal equality within

the US legal framework as a "blind" language model. This means that when a model is prompted to choose a race to fill in a blank, it assigns an equal probability to all options. They proceeded to fine-tune a pre-trained GPT-2 model using various subsets of American case law texts. This fine-tuning process accounted for the temporal context and the political leanings of the state hosting each court. The authors aimed to identify and quantify any unfairness in these models, specifically measuring the extent to which they deviated from legal equality by potentially associating different races with different legal contexts.

3.2 Data

3.2.1 Multilingual legal text data. The data found in the legal field are distinguished by their richness, diversity and professionalism. Large high-quality data sets are crucial for training models. Therefore, when building a model in the legal field, its performance greatly depends on the quality of the input data. As shown in Figure 4, this paper reviews the research efforts related to the collection of data in the legal field. Trklja and McAuliffe (2018) gathered and published a corpus consisting of legal text in multiple languages acquired from various jurisdictions. The corpus encompasses a diverse range of topics, such as the historical evolution processes of the meanings of legal terms, the migration of terms between legal systems, the use of binominals and the distribution of formulaic expressions in EU legal sublanguages. The primary objective of the corpus is to contribute to the ongoing development of the emerging field of language and law.

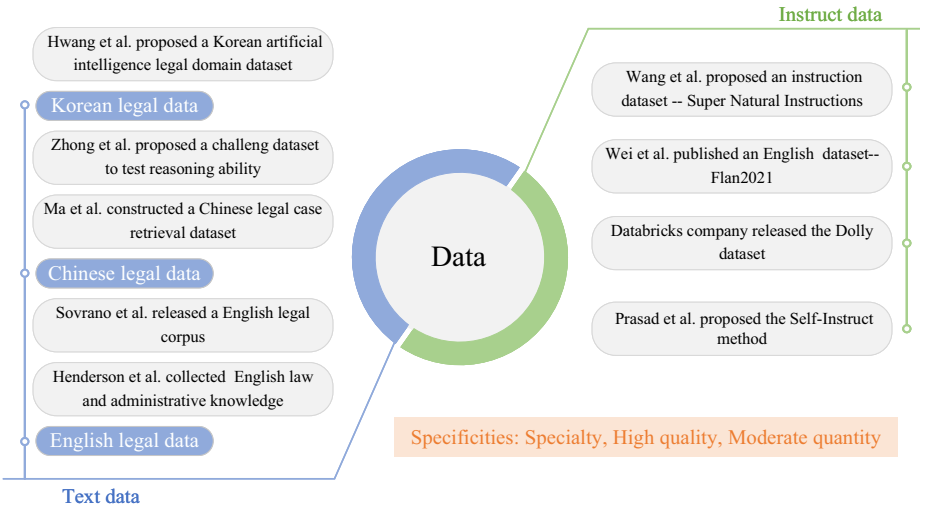
3.2.1.1 Korean legal data. Hwang *et al.* (2022) introduced the first large-scale benchmark dataset for the Korean legal AI domain. This data set comprises a legal domain corpus, two classification tasks, two legal judgment prediction (LJP) tasks and a summary task. The legal corpus encompasses numerous Korean legal cases, approximately half of which stem from cases adjudicated in the past four years. These cases involve multiple factual issues and have been reviewed by first- and second-level courts. The classification tasks involve categorizing case names and statutes based on factual descriptions of the cases. For the judgment prediction task, the samples include criminal and civil cases. The constructed model takes facts as inputs and produces an output indicating whether labor imprisonment will occur. The summary task contains supreme court cases and their corresponding summaries.

3.2.1.2 Chinese legal data. Zhong *et al.* (2020b) proposed a data set on which it is extremely challenging for a model to achieve high performance. They collected a portion of the data from the National Judicial Examination to construct a Q&A data set in the Chinese

Table 2. Overview and comparison of the developed model tuning methods

Methods	Advantages	Disadvantages	Application scenarios
Fine-tuning	Simplicity; flexibility; performance improvement	Costly computations	Sufficient data and resources; most NLP tasks
Soft prompt tuning	Does not change the model architecture	Prompts must be carefully designed	Limited resources scenarios
Adapter tuning	Efficiency; low resource	Limited effectiveness	Resource-constrained scenarios; cases in which a model must quickly adapt to new tasks
LoRA	Introducing low-rank matrices; reducing the number of required parameters	Limited effectiveness	Scenarios requiring efficient parameter adjustment

Source: Authors' creation



Source: Author's Creation

Figure 4.
Complete view of the works conducted on legal data collection

legal field. The characteristics of this data set are reflected in its test of the reasoning ability of a model. As the retrieval of relevant materials for Q&A tasks requires logical reasoning capabilities, such model proficiencies can be effectively tested on this dataset. The experiments conducted by the authors showed that even the state-of-the-art models could achieve an accuracy of only approximately 28% on this data set, while professional lawyers and untrained humans could achieve accuracies of 81% and 64%, respectively. Ma *et al.* (2021) constructed a Chinese legal case retrieval dataset that contains a small number of query cases and many candidate cases. The case data comes from criminal cases published by the Supreme People's Court of China. Owing to the difficulty of defining relevance, the authors proposed a series of relevance judgment criteria designed by a legal team, and legal experts annotated the corresponding candidate cases. This data set has important reference value for evaluating the retrieval performance of a model.

3.2.1.3 English legal data. Sovrano *et al.* (2021) released an English legal corpus and a legal knowledge exploration benchmark to facilitate the evaluation of automated Q&A methods in international private law settings. This benchmark enables detailed analyses of the performance of automated Q&A models developed for legal purposes. When studying the interactions between the original goals, the acquired knowledge, and the legal language understanding capabilities of a model, the sizes of the model and the training corpus are considered important evaluation dimensions. Henderson *et al.* (2022) collected a data set containing a large amount of English legal and administrative knowledge. It contains court opinions, contracts, administrative rules and legislative records. A law-based information filtering method was proposed; this approach effectively solves the problem that the data used by a model during the training process may contain biased, copyright-infringing and private information.

3.2.2 Instructional data tuning. In addition to equipping a model with knowledge from the legal field, it is important to enable the model to correctly understand human needs in a natural language format. The instructional data needed for the fine-tuning phase primarily

consists of natural language instances. Following the data collection step, the model undergoes fine-tuning in a supervised manner. The quality of the input instructional data directly impacts the final supervised fine-tuning effect. To enhance the performance of law LLMs, the data collected during the fine-tuning phase should possess the following characteristics.

3.2.2.1 Data source specialty. The instructional data used for data construction can be divided into manual and automatic data. The method of manually constructing instructions is relatively intuitive. Large amounts of Q&A data acquired from the internet can be filtered, or prompts and their corresponding answers written by annotators can be used. Although this is a relatively labor-intensive process, it can effectively control the instructional data annotation procedure and the overall quality of these data. However, manually constructing instructional data is expensive and requires considerable manpower. Therefore, some studies have attempted to find more efficient alternatives. Prasad *et al.* (2023) proposed a gradient-free instructional prompt search method. This is a gradient-free, edit-based search approach that enhances task instructions for LLMs. Human-designed instructions are used as inputs, and automatically, an improved, edited prompt is provided, allowing for API-based tuning.

3.2.2.2 High data quality requirements and moderate quantity requirements. The quantity and quality of instructional data significantly influence the performance of the developed model. Specifically, this is reflected in the following aspects. First, using a large amount of instructional data can improve the generalization capabilities of LLMs. However, as the number of tasks increases, the model performance improvement provided becomes minimal. Second, a certain number of instruction data sets can provide relatively sufficient knowledge, while the use of redundant data may result in overfitting. Zhou *et al.* (2023) showed that high-quality and diverse instructional data can yield better results. When using only 1,000 high-quality, rich and diverse instructional data points, models can reach or exceed the performance of other models with the same parameters but using dozens of times more data points. Specifically, the authors mainly performed the following work in terms of data quality assurance. First, the top-ranked high-quality answers were selected, and all the answers were unified according to the style of an AI assistant. Second, answers that were too long or too short, began in the first person or contained links were deleted. Third, carefully hand-coded answers were chosen by annotators.

We review some of the related research on the construction of instructional data for tuning. Wang *et al.* (2022) proposed an instruction dataset called Super Natural Instructions. The data set consists of 1,600 NLP tasks in 55 languages. Each task consists of two parts: "instructions" and "task examples". The "instructions" section not only describes each task in detail but also provides positive and negative examples and their corresponding explanations. The "task examples" are input-output instances belonging to the target task. Wei *et al.* (2021) published an English instruction data set called Flan2021, which was constructed by converting 62 widely used NLP benchmarks into input–output pairs. During the data set formation process, first, directives and target templates were manually written, and then the templates were populated with data instances from the data set. The Databricks company released the Dolly (2023) data set. This data set contains 15,000 artificially constructed English instructional data points, aiming to simulate a wide range of human behaviors to promote ChatGPT-like interactivity in LLMs. The survey included seven task types: open-question answering, closed-question answering, information extraction, summarization, brainstorming, classification and creative writing.

Table 3 presents a comprehensive comparative analysis among various data collection studies.

Works	Content	Characteristic description	Purpose
Trklja and McAuliffe (2018)	Legal case corpus	Multi-language	Promote research for legal field
Hwang *et al.* (2022)	Legal case corpus; classification task; judgment prediction task; summary task	Korean; latest content	A legal model benchmark
Zhong *et al.* (2020b)	China's National Judicial Examination	Test the legal reasoning capabilities of AI models	Evaluating Chinese legal AI models
Sovrano *et al.* (2021)	English legal text corpus; court opinions; contracts; administrative rules;	Proposing new information filtering methods to protect privacy	Evaluate model performance
Wang and Tian (2023)	Instruction tuning data	High-quality data	Providing relatively sufficient knowledge for models; make models more interactive

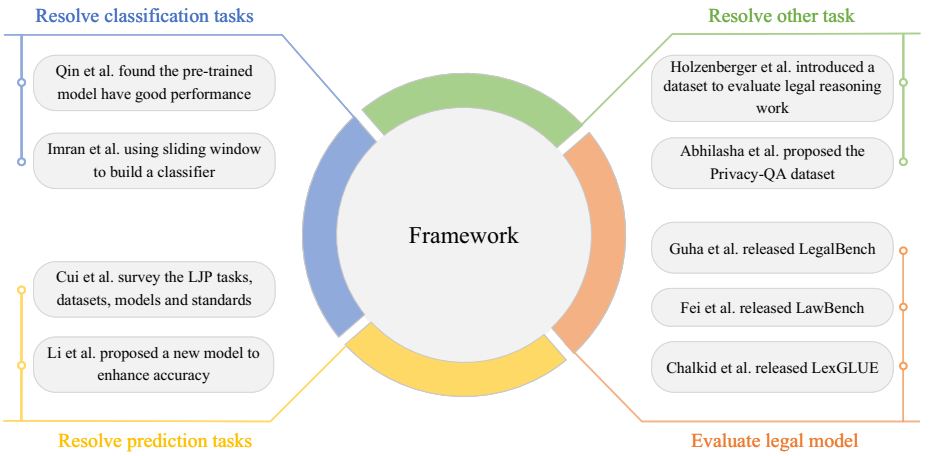
Source: Authors' creation

Table 3. Existing data collection works involving legal AI and LLMs

3.3 Frameworks

3.3.1 Frameworks for resolving different legal tasks. Addressing distinct legal tasks necessitates the use of varying legal data and task frameworks for support. Many different studies have also targeted different legal tasks. Figure 5 categorizes some of the related research from the perspective of solving legal tasks.

3.3.1.1 Legal document classification task. The classification of legal documents is the basis of legal artificial intelligence tasks and has important research value. Qin *et al.* (2022) found that, compared with machine learning models based on feature engineering and traditional convolutional neural network or recurrent neural network models in the field of NLP, language models pre-trained on an English corpus achieve good performance in classification tasks. Several different pre-trained language models have been studied, and the Chinese legal corpus has been used for pretraining. This approach is highly efficient for



Source: Author's creation

Figure 5. Complete overview of the frameworks used to address legal issues

classification tasks in all legal subfields. In the legal field, limited works have used pre-trained language models to classify long documents. Imran *et al.* (2023) proposed a legal document classifier that classifies legal documents by using a sliding window method. This method effectively increases the maximum sequence length that a model can handle.

3.3.1.2 Legal judgment prediction task. LJP is the application of NLP technology to automatically predict judgment results based on factual descriptions. Cui *et al.* (2023a) conducted a comprehensive survey on the existing LJP tasks, data sets, models and evaluation standards. First, LJP data sets constructed in different languages were analyzed and summarized into LJP classification methods based on three different attributes. Second, the used evaluation indicators were summarized and divided into different types to evaluate the performance attained by an LJP model in different tasks. Third, pre-trained models developed for the legal field in different languages were studied, and the LJP task was highlighted. Li *et al.* (2020) proposed a new charge prediction model through criminal justice interpretation to provide more accurate charge predictions. This model introduced the crime explanation concept, which alleviates the problems caused by data imbalances and confusing charges. Through embedding techniques, both factual descriptions and crime explanations were embedded into low-dimensional vector spaces and neural networks, providing achievable, computable charge predictions.

3.3.1.3 Other law tasks. Statutes can be viewed as collections of normative rules expressed in natural language, and the process of applying statutes to the facts of a case is called statutory reasoning. Holzenberger *et al.* (2020) introduced a text corpus data set in the legal field to evaluate legal reasoning work. This task differs from most traditional machine reading tasks in that the information necessary for producing a judgment must be accurately presented.

Furthermore, privacy policies involve long and complex documents that are often difficult for users to read and understand. Ravichander and Alan (2019) proposed the PrivacyQA data set, which consists of a complex corpus composed of multiple questions concerning mobile application privacy policies, as well as expert annotations for the related answers. This dataset has implications regarding the study of legal models for privacy tasks. It provides a reference for the assessment of how applications can legally collect, manage and use user data.

3.3.2 Frameworks for evaluating legal models. Owing to the rapid development trends of LLMs, LLM applications have demonstrated strong potential from various perspectives. However, when such models are applied to the highly specialized field of law, it is unclear how much legal knowledge the models can possess and whether they can reliably perform law-related tasks. The existing NLP task benchmarks can no longer satisfy the evaluation needs of LLMs.

Rasiah *et al.* (2023) proposed a new NLP task evaluation benchmark. This benchmark challenges the current LLMs in four key aspects: processing long documents, leveraging domain-specific knowledge, performing multilingual understanding and multitasking. Additionally, diverse NLP data sets in the legal field have been collected from various national legal systems, enabling a comprehensive study of multilingual legal models.

Guha *et al.* (2023) proposed LegalBench, which is a benchmark test set for legal reasoning-based LLMs that covers six types of legal reasoning tasks. The benchmark test set was constructed by using an interdisciplinary research process, and it contains tasks designed and produced by legal professionals. It mainly includes tasks that can measure actual legal reasoning capabilities or reasoning tasks that are of interest to lawyers.

Fei *et al.* (2023) proposed LawBench, which is an evaluation framework that was specifically designed to evaluate the performance of LLMs in legal tasks. This framework

accurately evaluates the legal capabilities of LLMs at three cognitive levels. The first factor is legal knowledge memory, which refers to whether LLMs can remember the required legal documents and factual information. The second factor concerns legal knowledge understanding, i.e. whether LLMs can understand the entities, events and relationships contained in legal texts. The third factor involves the application of legal knowledge, that is, whether LLMs can correctly apply legal knowledge and perform the necessary reasoning steps to solve real-world legal problems.

Chalkidis *et al.* (2022) proposed the Legal General Language Understanding Evaluation Benchmark, which is a collection of data sets used to evaluate the performance of models in various legal language-based understanding tasks in a standardized manner. The core of this evaluation method is to successfully quantify the legal text analysis performance of a model by using natural language understanding technology. This approach is an important analysis and evaluation tool for legal workers. By using this method to evaluate model performance, the result is an appropriate model for solving problems concerning legal text. The ever-increasing scale of text analysis has made the related tasks increasingly complex.

4. Challenges and prospects

4.1 Prospects for law-based large language model applications

The legal field is built upon specialized knowledge, thus forming a professional domain. Recently, studies have used LLMs to address various legal tasks, such as legal document analysis, LJP and legal consultation Q&A tasks. These studies have revealed that LLMs exhibit significant potential for extensive applications within the legal field.

4.1.1 Legal document research. LLMs showcase remarkable proficiency in terms of processing textual data. This high level of proficiency proves beneficial for legal professionals, enabling them to swiftly extract valuable information from extensive textual data sets. Sasidharan and Rahulnath (2023) proposed a structured approach that combines relation extraction, graph reasoning and clustering methods to effectively extract and represent legal relationships. First, they used NLP and machine learning algorithms to recognize entity pairs and assign appropriate relationship labels. Next, they leveraged graph reasoning to enhance the extracted relationships with additional contextual information by representing legal entities and their relationships as nodes and edges, respectively, in a knowledge graph and applying inference algorithms to infer implicit relationships and expand the graph according to rules and regulations in the legal domain. This approach effectively facilitates decision-making and knowledge discovery in the legal field. Additionally, LLMs have the potential to offer advantages for constructing knowledge graphs and updating legal knowledge bases. Overall, LLMs contribute to the deepening and enhancement of legal research.

4.1.2 Case prediction analysis. Owing to the complexity and diversity of legal cases, it is essential to furnish personalized decision-making solutions that are tailored to the specific needs of relevant parties. LJP involves automatically predicting the judgment results of a legal case based on its factual description, and this task has excellent prospects in judicial assistance systems and consultation services for the public. He *et al.* (2023a) introduced legal charge knowledge to enhance the representations of fact descriptions and incorporated it into deep neural networks. Then, a cross-attention mechanism was proposed to model the relationship between legal charge knowledge and fact descriptions in a unified model. This approach improves the performance achieved in LJP tasks and solves legal cases involving confusing charges and legal terms. LLMs possess prediction and classification capabilities; when paired with parties' case information and a wealth of legal data, they facilitate the delivery of customized legal advice to lawyers or other involved parties. This personalized

advice assists lawyers and other parties in reviewing cases from a comprehensive perspective.

4.1.3 Legal consultation-based question-and-answer tasks. By leveraging NLP technology, legal consultation systems can answer questions posed by users and provide accurate explanations and answers concerning legal compliance. Tamatjita *et al.* (2021) used predicate calculus and Bayes' theorem to construct decision rules and applied Iterative Dichotomiser 3 as the production rule to calculate the information gain derived from a decision tree. Based on this, they built an expert system for legal consultations regarding song royalties to help song creators make decisions. Wang *et al.* (2020) designed a legal consultation system for marriage cases based on a parallel C4.5 decision tree. First, they converted legal consultation problems into classification tasks. Second, a legal consultation classification prediction model based on the parallel C4.5 decision tree algorithm was trained with MapReduce using the collected data. Finally, a model based on the SVM algorithm, possessing a strategy that was designed to provide automatic interactions for users, was developed. LLMs have strong capabilities for solving such tasks. Users can seek guidance to address questions related to specific industry regulations or national standards, and LLMs can provide compliance advice and guidance based on the corresponding legal documents or best-practice cases. This approach helps users adhere to legal documents and resolve any uncertainties they may have.

4.2 Case studies regarding law-based large language models applications

4.2.1 ChatLaw. Cui *et al.* (2023b) proposed an open-source LLM called ChatLaw. This is a Chinese law LLM with LLaMA-13B serving as its base model. The research related to this model was primarily characterized by the following key aspects. First, the authors created a fine-tuned data set for the legal field. This data set was constructed from an extensive corpus of textual data derived from the legal field and encompasses diverse sources, such as legal judgment documents, judicial interpretations and legal examination questions. Second, to address the model hallucination issue encountered during the reference retrieval and legal data screening processes, judgment cases were used to construct a training data set. A similarity-matching model was trained on this data set to correlate user-provided case information with the pertinent corresponding legal references. Third, a methodology that integrates vector database retrieval with the keyword search process was introduced, where the goal was to mitigate the inaccuracies associated with relying solely on vector database retrieval. Fourth, a self-attention method was introduced to enhance the self-checking capabilities of the LLM. This approach aimed to prevent a decrease in the accuracy of the model output resulting from errors in the reference data. Based on the above work, this research contributed a pragmatic solution for the integration of LLMs and knowledge bases within legal scenarios.

4.2.2 DISC-LawLLM. Yue *et al.* (2023) released a law LLM called DISC-LawLLM. The objective of this model is to construct an AI-based legal system with reasoning and knowledge retrieval capabilities. Specifically, in terms of legal document processing, the model necessitates the ability to comprehend and generate professional sentences. Therefore, training data was constructed by leveraging existing publicly available judicial data and legal document data. In terms of reasoning capabilities, the model must possess legal reasoning proficiency. Therefore, some of the data were designed based on the reasoning theory of legal syllogism. To enable the model to accurately retrieve legal documents, the authors incorporated a retrieval enhancement module. This approach enhanced the proficiency of the model in terms of retrieving background knowledge and adhering to legal documents. To evaluate their model, they developed an evaluation

framework. The evaluation assessed the performance of the model from objective and subjective viewpoints. During the objective evaluation, the data set incorporated a series of single-item and multiple-choice questions sourced from China's standardized legal examinations and knowledge competitions. On the subjective side, they used Q&A formats to simulate the process of administering subjective examination questions.

4.2.3 Case summary. The amount of research concerning legal LLMs has increased in recent years. The related studies primarily focused on the selection of parameters, pretraining data, fine-tuning data and base models. First, the developed models typically boast billions of parameters, enabling them to excel at understanding and processing complex legal texts. Second, the quality and diversity of the used pretraining data are paramount for maximizing model performance. In the legal domain, such data often encompasses legal documents, cases and statutes, among others, providing rich background knowledge to the constructed model. Third, the selection and quality of the fine-tuning data directly impact the performance of a model in specific tasks. During fine-tuning, the model is fed domain-specific input data and adjusted by optimizing the objective function to adapt to particular legal application scenarios. Finally, the chosen base model is also a critical factor that affects model performance. Different base models may have distinct architectures and characteristics, thereby influencing their performance on various tasks. Table 4 presents a comprehensive comparative analysis of various legal-specific LLMs.

4.3 Challenges faced by law-based large language model applications

4.3.1 Accuracy of the generated content. LLMs have the potential to generate content that appears plausible but is factually incorrect. McIntosh *et al.* (2023) noted that LLMs hallucinate, which refers to the generation of misinformation or outputs that are not aligned with reality. LLMs are frequently black-box models, making it challenging to interpret their internal decision-making processes. This poses a significant challenge in the legal field, as lawyers and clients require comprehension of the legal document basis and reasoning process behind the decisions of a model to understand and accept the recommendations or outcomes provided by the model. Countermeasures for addressing this challenge involve conducting interpretability research to explore methods that can visualize or explain the decision-making processes of LLMs to lawyers and clients. This includes the development of interpretive technologies and methods tailored to the legal field, which aims to render the decision-making processes of

Model name	Number of parameters	Pretraining data	Instruction data	Base model
LawGPT[a]	7 B	✓	✓	Chinese-LLaMA
LexiLaw[b]	6 B	✓	✓	ChatGLM
Wisdom-Interrogatory[c]	7 B	✓	✓	Baichuan
HanFei (He *et al.*, 2023b)	7 B	✘	✓	HanFei
DISC-LawLLM (Yue *et al.*, 2023)	13 B	✘	✓	Baichuan
Fuzi-Mingcha (Wu *et al.*, 2023)	6 B	✓	✓	ChatGLM
Lawyer-LLaMA (Huang *et al.*, 2023b)	13 B	✘	✓	LLaMA
ChatLaw (Cui *et al.*, 2023c)	13/33 B	✘	✘	Ziya-LLaMA/ Anima

Notes: [a]https://github.com/pengxiao-song/LaWGPT; [b]https://github.com/CSHaitao/LexiLaw; [c]https://github.com/zhihaiLLM/wisdomInterrogatory
Source: Authors' creation

Table 4. Legal-specific LLMs

models more transparent, understandable and acceptable. Additionally, providing references that cite the legal bases of model decisions ensures that lawyers and clients can comprehend and appropriately utilize the recommendations provided by the model.

4.3.2 Lack of a high-quality legal corpus. The quality of a legal document corpus includes both the quality of the original data and the quality of its corresponding labels (Chen *et al.*, 2022). The major quality dimensions of a legal document corpus include comprehensiveness, freshness and correctness. The absence of a comprehensive, correct and fresh legal document corpus presents a notable challenge for law LLMs. Compared with general documents, legal documents are usually long text sequences with inherent logical structures (Li *et al.*, 2023). However, most existing language models have difficulty understanding the long-distance dependencies among different structures. Moreover, compared with general retrieval, legal case retrieval is more sensitive to key legal elements. Even subtle differences in the elements of key laws can affect the judgments produced for related cases. However, the existing pre-trained language models designed for general purposes have not been equipped to handle legal elements. The construction of a high-quality corpus requires professional legal knowledge and the participation of lawyers, and the construction process is complex and time-consuming. Countermeasures for solving this problem include establishing standardized data collection and preprocessing strategies that can remove data noise and correct data errors. A multidisciplinary collaborative team was established to combine the knowledge and professional judgments of legal experts to improve the quality and accuracy of the constructed data.

4.3.3 Security and privacy issues. The security and privacy concerns encountered when deploying LLMs primarily encompass issues such as data leakage, unauthorized data access and data abuse. Legal data contains sensitive information about the involved parties, including their home addresses, their mobile phone numbers and other personally identifiable information. Consequently, it is imperative to implement effective measures that can ensure the security of such data. Countermeasures for addressing this issue include the implementation of data encryption technology that can safeguard the transmission and storage of data. Another essential countermeasure is the establishment of stringent data access and authority control mechanisms, ensuring that only authorized personnel have access to sensitive data. In addition, it is crucial to design and enforce appropriate data-sharing policies and contracts. These rules should explicitly outline the purposes and limitations of how data may be used, ensuring their alignment with the relevant legal documents an0d regulations. Strengthening the current monitoring and auditing processes for data security is essential, as they enable the prompt detection of security vulnerabilities or threats and quick responses.

5. Summary

The advent of LLMs represents an emerging paradigm for addressing the construction requirements of intelligent justice systems. In the future, law LLMs are poised to evolve into crucial tools and resources for resolving legal issues. These models have the potential to offer lawyers and clients more precise and personalized legal retrieval and assistance services. First, this paper delves into an extensive examination of LLMs, introducing their concepts and tracing their developmental history. Furthermore, this paper surveys the most recent advancements in law LLM research. Second, this paper consolidates insights into diverse facets for fine-tuning LLMs in the legal field. It encompasses discussions on PEFT algorithms, legal domain data and frameworks tailored for legal tasks. Additionally, this paper offers a comprehensive review of the potential applications of LLMs in the legal domain, with an examination of specific legal LLM application instances. Finally, this paper highlights the key challenges faced by law LLMs. The work we perform in this paper can

help not only researchers but also legal practitioners by continuously applying research and legal practice to explore the potential of LLMs in the legal field. Future research can achieve further breakthroughs and innovations and contribute to the legal and AI application fields, thus benefiting legal practitioners and legal communities worldwide by helping them access more humanized, intelligent and precise legal services.

References

Aghajanyan, A., Gupta, S. and Zettlemoyer, L. (2021), "Intrinsic dimensionality explains the effectiveness of language model fine-tuning", Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7319-7328.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.* (2020), "Language models are few-shot learners", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877-1901.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. and Aletras, N. (2022), "Lexglue: a benchmark dataset for legal language understanding in English", Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4310-4330.

Chen, H., Pieptea, L.F. and Ding, J. (2022), "Construction and evaluation of a high-quality corpus for legal intelligence using semiautomated approaches", *IEEE Transactions on Reliability*, Vol. 71 No. 2, pp. 657-673.

Cui, J., Shen, X. and Wen, S. (2023a), "A survey on legal judgment prediction: datasets, metrics, models and challenges", *IEEE Access*, Vol. 11.

Cui, J., Li, Z., Yan, Y., Chen, B. and Yuan, L. (2023b), "Chatlaw: open-source legal large language model with integrated external knowledge bases", arXiv preprint arXiv:2306.16092.

Cui, J., Li, Z., Yan, Y., Chen, B. and Yuan, L. (2023c), "Chatlaw", available at: https://github.com/PKU-YuanGroup/ChatLaw

Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L. (2023), "Qlora: efficient finetuning of quantized LLMS", arXiv preprint arXiv:2305.14314.

Dias, J., Santos, P.A., Cordeiro, N., Antunes, A., Martins, B., Baptista, J. and Gonçalves, C. (2022), "State of the art in artificial intelligence applied to the legal domain", arXiv preprint arXiv:2204.07047.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., *et al.* (2023), "Parameter-efficient fine-tuning of large-scale pre-trained language models", *Nature Machine Intelligence*, Vol. 5 No. 3, pp. 220-235.

Dolly, F. (2023), "Introducing the world's first truly open instruction-tuned LLM", available at: databricks.com

El Moussaoui, T., Chakir, L. and Boumhidi, J. (2023), "Preserving privacy in Arabic judgments: Ai-powered anonymization for enhanced legal data privacy", *IEEE Access*, Vol. 11.

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z. and Ge, J. (2023),: "Lawbench: benchmarking legal knowledge of large language models", arXiv preprint arXiv:2309.16289.

Fernandes, B.J., Cleyton, M.D.O., Silva, L.H., Rocha, A.L., Sobral, P.C., *et al.* (2023), "Fine-tuning multifit for enhanced legal sentence basis classification", *IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, *IEEE*, pp. 1-6.

Gordon, D.G. and Breaux, T.D. (2014), "The role of legal expertise in interpretation of legal requirements and definitions", *IEEE 22nd International Requirements Engineering Conference (RE)*, *IEEE*, pp. 273-282.

Guha, N., Nyarko, J., Ho, D.E., Re, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., *et al.* (2023), "Legalbench: a collaboratively built benchmark for measuring

legal reasoning in large language models", Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Hauser, M.D., Chomsky, N. and Fitch, W.T. (2002), "The faculty of language: what is it, who has it, and how did it evolve?", *Science*, Vol. 298 No. 5598, pp. 1569-1579.

He, C., Tan, T.P., Zhang, X. and Xue, S. (2023a), "Knowledge-enriched multi-cross attention network for legal judgment prediction", *IEEE Access*, Vol. 11.

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T. and Neubig, G. (2021), "Towards a unified view of parameter-efficient transfer learning", International Conference on Learning Representations.

He, S., Ding, L., Dong, D., Zhang, J. and Tao, D. (2022), "Sparseadapter: an easy approach for improving the parameter-efficiency of adapters", Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 2184-2190.

He, W., Wen, J., Zhang, L., Cheng, H., Qin, B., Li, Y., Jiang, F., Chen, J., Wang, B. and Yang, M. (2023b), Hanfei-1.0, available at: https://github.com/siat-nlp/HanFei

Henderson, P., Krass, M., Zheng, L., Guha, N., Manning, C.D., Jurafsky, D. and Ho, D. (2022), "Pile of law: learning responsible data filtering from the law and a 256gb open-source legal dataset", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 29217-29234.

Holzenberger, N., Blair-Stanek, A. and Van Durme, B. (2020), "A dataset for statutory reasoning in tax law entailment and question answering".

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M. and Gelly, S. (2019), "Parameter-efficient transfer learning for NLP", *International Conference on Machine Learning*, pp. 2790-2799, *PMLR*.

Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., *et al.* (2021), "Lora: low-rank adaptation of large language models", International Conference on Learning Representations.

Huang, Q., Tao, M., An, Z., Zhang, C., Jiang, C., Chen, Z., Wu, Z. and Feng, Y. (2023a), "Lawyer llama technical report", arXiv preprint arXiv:2305.15062.

Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z. and Feng, Y. (2023b), "Lawyer llama", available at: https://github.com/AndrewZhe/lawyer-llama

Hwang, W., Lee, D., Cho, K., Lee, H. and Seo, M. (2022), "A multi-task benchmark for Korean legal language understanding and judgement prediction", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 32537-32551.

Imran, A.S., Hodnefjeld, H., Kastrati, Z., Fatima, N., Daudpota, S.M. and Wani, M.A. (2023), "Classifying European court of human rights cases using transformer-based techniques", *IEEE Access*, Vol. 11.

Karimi, M.R., Henderson, J. and Ruder, S. (2021), "Compacter: Efficient low-rank hypercomplex adapter layers", *Advances in Neural Information Processing Systems*, Vol. 34, pp. 1022-1035.

Katz, D.M., Hartung, D., Gerlach, L., Jana, A. and Bommarito, M.J. (2023), "Natural language processing in the legal domain", available at SSRN 4336224.

Lai, J., Gan, W., Wu, J., Qi, Z. and Yu, P.S. (2023), "Large language models in law: a survey", arXiv preprint arXiv:2312.03718.

Lee, J.S. (2023), "LexGPT 0.1: pre-trained GPT-j models with pile of law", arXiv preprint arXiv:2306.05431.

Lester, B., Al-Rfou, R. and Constant, N. (2021), "The power of scale for parameter-efficient prompt tuning", Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045-3059.

Li, X.L. and Liang, P. (2021), "Prefix-tuning: optimizing continuous prompts for generation", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582-4597.

Li, X., Kang, X., Wang, C., Dong, L., Yao, H. and Li, S. (2020), "A neural-network-based model of charge prediction via the judicial interpretation of crimes", *IEEE Access*, Vol. 8, pp. 101569-101579.

Li, H., Ai, Q., Chen, J., Dong, Q., Wu, Y., Liu, Y., Chen, C. and Tian, Q. (2023), "Sailer: structure-aware pre-trained language model for legal case retrieval", arXiv preprint arXiv:2304.11370.

McIntosh, T.R., Liu, T., Susnjak, T., Watters, P., Ng, A. and Halgamuge, M.N. (2023), "A culturally sensitive test to evaluate nuanced GPT hallucination", *IEEE Transactions on Artificial Intelligence*, Vol. 1 No. 1, pp. 1-13.

Ma, L. (2022), "Artificial intelligence in legal education under the background of big data computation", *International Conference on Computation, Big-Data and Engineering (ICCBE)*, pp. 51-53. *IEEE*.

Ma, Y., Shao, Y., Wu, Y., Liu, Y., Zhang, R., Zhang, M. and Ma, S. (2021), "Lecard: a legal case retrieval dataset for Chinese law system", *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2342-2348.

Malic, V.Q., Kumari, A. and Liu, X. (2023), "Racial skew in fine-tuned legal AI language models", IEEE International Conference on Data Mining Workshops (ICDMW), pp. 245-252.

Nguyen, H.T. (2023), "A brief report on LAWGPT 1.0: a virtual legal assistant based on gpt-3", arXiv preprint arXiv:2302.05729.

OpenAI, R. (2023), "Gpt-4 technical report. arxiv 2303.08774", View in Article.

Prasad, A., Hase, P., Zhou, X. and Bansal, M. (2023), "Grips: gradient-free, edit-based instruction search for prompting large language models", *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3827-3846.

Qin, R., Huang, M. and Luo, Y. (2022), "A comparison study of pre-trained language models for Chinese legal document classification", 5th International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE, pp. 444-449.

Qu, Y., Zhang, Z. and Bai, B. (2023), "The way forward for legal knowledge engineers in the big data era with the impact of ai technology", 6th International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE, pp. 225-229.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., *et al.* (2018), "Improving language understanding by generative pre-training".

Rasiah, V., Stern, R., Matoshi, V., Stürmer, M., Chalkidis, I., Ho, D.E. and Niklaus, J. (2023), "Scale: scaling up the complexity for advanced language model evaluation", arXiv preprint arXiv:2306.09237.

Ravichander, A. and Alan, W. (2019), "Question answering for privacy policies: combining computational and legal perspectives", *Empirical Methods in Natural Language Processing*.

Resck, L.E., Ponciano, J.R., Nonato, L.G. and Poco, J. (2022), "Legalvis: exploring and inferring precedent citations in legal documents", *IEEE Transactions on Visualization and Computer Graphics*.

Sabry, M. and Belz, A. (2023), "Peft-ref: a modular reference architecture and typology for parameter-efficient finetuning techniques", arXiv preprint arXiv:2304.12410.

Sasidharan, A.K. and Rahulnath, R. (2023), "Structured approach for relation extraction in legal documents", 4th IEEE Global Conference for Advancement in Technology (GCAT), IEEE, pp. 1-6.

Sheik, R. and Nirmala, S.J. (2021), "Deep learning techniques for legal text summarization", IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), IEEE, pp. 1-5.

Song, D., Gao, S., He, B. and Schilder, F. (2022), "On the effectiveness of pre-trained language models for legal natural language processing: an empirical study", *IEEE Access*, Vol. 10, pp. 75835-75858.

Sovrano, F., Palmirani, M., Distefano, B., Sapienza, S. and Vitali, F. (2021), "A dataset for evaluating legal question answering on private international law", *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 230-234.

Su, Y., Wang, X., Qin, Y., Chan, C.M., Lin, Y., Wang, H., Wen, K., Liu, Z., Li, P., Li, J., *et al.* (2022), "On transferability of prompt tuning for natural language processing", Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3949-3969.

Tamatjita, E.N., Sobalely, A.P., Budiyantara, A. and Silitonga, A.V. (2021), "Expert system for legal consultation of song royalty with iterative dichotomiser 3 algorithm", 6th International Conference on New Media Studies (CONMEDIA), IEEE, pp. 84-89.

Taulli, T. (2023), "Large language models: how generative AI understands language", *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business*, Springer, pp. 93-125.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E. and Azhar, F. (2023a), "Llama: open and efficient foundation language models", arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., *et al.* (2023b), "Llama 2: open foundation and fine-tuned chat models", arXiv preprint arXiv:2307.09288.

Trklja, A. and McAuliffe, K. (2018), "The European union case law corpus (euclcorp): a multilingual parallel and comparative corpus of EU court judgments", in Frank, A.U., Ivanovic, C., Mambrini, F., Passarotti, M. and Sporleder, C. (Eds), *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities: CRH-2 Gerastree Proceedings*, Vol. 1, pp. 217-226.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. and Kaiser, P. (2017), "I.: attention is all you need", *Advances in Neural Information Processing Systems*, Vol. 30.

Wagh, R. and Anand, D. (2017), "Application of citation network analysis for improved similarity index estimation of legal case documents: a study", IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), pp. 1-5.

Wang, Z. (2021), "China's e-justice revolution", *Judicature*, Vol. 105, p. 36.

Wang, N. and Tian, M.Y. (2023), "'Intelligent justice': human-centered considerations in China's legal AI transformation", *AI and Ethics*, Vol. 3 No. 2, pp. 349-354.

Wang, C., Chen, D., Hu, Y., Ceng, Y., Chen, J. and Li, H. (2020), "Automatic dialogue system of marriage law based on the parallel c4. 5 decision tree", *IEEE Access*, Vol. 8, pp. 36061-36069.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R. (2018), "Glue: a multi-task benchmark and analysis platform for natural language understanding", International Conference on Learning Representations.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A.S., Arunkumar, A. and Stap, D. (2022), "Super-naturalinstructions: generalization via declarative instructions on 1600+ NLP tasks", *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085-5109.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.* (2022), "Chain-of-thought prompting elicits reasoning in large language models", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24824-24837.

Wei, F., Keeling, R., Huber-Fliflet, N., Zhang, J., Dabrowski, A., Yang, J., Mao, Q. and Qin, H. (2023), "Empirical study of LLM fine-tuning for text classification in legal document review", IEEE International Conference on Big Data (BigData), pp. 2786-2792, *IEEE*.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N. and Dai, A.M. (2021), "Q.V.: finetuned language models are zero-shot learners", International Conference on Learning Representations.

Wu, S., Liu, Z., Zhang, Z., Chen, Z., Deng, W., Zhang, W., Yang, J., Yao, Z., Lyu, Y., Xin, X., Gao, S., Ren, P., Ren, Z. and Chen, Z. (2023), fuzi.mingcha, available at: https://github.com/irlab-sdu/fuzi.mingcha

Yang, A., Xiao, B., Wang, B., Zhang, B., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., Yang, F., *et al.* (2023), "Baichuan 2: open large-scale language models", arXiv preprint arXiv:2309.10305.

Yue, S., Chen, W., Wang, S., Li, B., Shen, C., Liu, S., Zhou, Y., Xiao, Y., Yun, S., Lin, W., *et al.* (2023), "Disc-lawllm: fine-tuning large language models for intelligent legal services", arXiv preprint arXiv:2309.11325.

Zhang, L., Zhang, L., Shi, S., Chu, X. and Li, B. (2023), "Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning", arXiv preprint arXiv:2308.03303.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W. and Zhao, T. (2022), "Adaptive budg*et al*location for parameter-efficient fine-tuning", The Eleventh International Conference on Learning Representations.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z. and Sun, M. (2020a), "How does NLP benefit legal system: a summary of legal artificial intelligence", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5218-5230.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z. and Sun, M. (2020b), "JEC-QA: a legal-domain question answering dataset", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 No. 5, pp. 9701-9708.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P. and Yu, L. (2023), "Lima: Less is more for alignment", arXiv preprint arXiv:2305.11206.

Corresponding author
Xiaoxian Yang can be contacted at: xxyang@sspu.edu.cn