Exploring Transformer-based LLMs Models, Applications, and Challenges in Law: A Survey

Peter Jingzhou Lai[1], Ling Xia Liao[2], Jie Chen[3], Miao Zhang[1], Han-Chieh Chao[4,5]*

[1] Fujian Provincial Key University Laboratory of Cloud Computing and IoT, Quanzhou University of Information Engineering, China
[2] School of Artificial Intelligence, Guilin University of Aerospace University, China
[3] School of Electronic Engineering and Automation, Guilin University of Electronic Technology, China
[4] Department of Applied Informatics, Fo Guang University, Taiwan
[5] Department of Electrical Engineering, National Dong Hwa University, Taiwan
laijingzhou@qzuie.edu.cn, liaolx@guat.edu.cn, jiechen@mails.guet.edu.cn, zm@qzuie.edu.cn, hcc@fgu.edu.tw

Abstract

Transformer-based Large Language Models (LLMs) have rapidly advanced the state of natural language processing, offering powerful capabilities in understanding text, and automating drafting. Given that law is a text-driven domain, such models transform_workflows in legal practice. This survey provides a comprehensive overview of 80 publications on Transformer-based LLM classes and their datasets in the legal domain. It simplifies the categorization of these models into three groups: general Transformer models, legal (domain-specific) LLMs, and task-specific models. For general models, we examine the encoder-only, decoder-only, and encoder–decoder structure, as well as domain-specific and commercial legal LLMs. We also provide legal datasets for LLM training and evaluation and identify open challenges and research directions that enhance deployment. By reviewing current Transformer model applications and limitations, this survey provides a task-based view of legal LLM classes and performance, in order to better understand the application implications for natural language processing and legal researchers, and future directions for legal LLMs to become effective and reliable tools for client-centred service and judicial economy.

Keywords: Transformer, LLM, Legal LLM, LLM-based legal application, NLP

1 Introduction

Transformer-based Large Language Models (LLMs) are AI systems built on the Transformer architecture. Using self-attention, this architecture can model long sequences for word (or token) relationships, capturing context, semantics, and dependencies much better than earlier models such as Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN). When developers scale tokens and parameters to the billions, the model is classified as LLMs.

Given model capacities to understand universal language and the language-intensive domains, Transformer-based LLMs can process gigabytes of documents involving scientific papers, clinical notes, financial filings, or legal judgements, and be fine-tuned for other task-specific documents [1].

Transformer-based LLMs have automated workflows in sub-domains from the natural to social sciences. In biomedicine and healthcare, pre-trained LLMs can summarize clinical notes, process patient-doctor questions and answers, mine medical literature, support faster diagnosis, and provide drug discovery insights [2]. In academic writing and research, relevant LLMs can summarize, classify, and extract facts from journal articles, and generate literature review drafts, codes, and data-driven discoveries [3]. In finance and marketing, general LLMs have been trained on financial data for market intelligence and automated summarization of financial reports, earnings calls, and compliance documents, reducing workload in risk analysis, trend forecasts, and other reports in banking and insurance [4]. And in education, LLMs have powered tutoring systems to become personalized learning assistants in explaining concepts in multiple ways, improving equity-based learning [5].

Especially in law, legal LLMs have proliferated over the past few years. Encoder-only models such as LegalBERT [6] excel at retrieval, classification, and entity recognition; encoder–decoder models such as LegalT5 [7] enable summarization and legal question answering; and decoder-only families such as SaulLM [8] demonstrate strong generative and reasoning capabilities. Each model family leverages diverse legal corpora, from holding statements to merger agreements, from the US Securities Commission to EurLex; and fine-tuning strategies now range from span corruption to tune instruction. However, this upsurge in legal tasks and datasets come with taxonomies that are hard to reconcile.

Problem Statement. Despite repeated research attempts to categorize, state-of-the-art legal LLMs still lack classes for model development [9-13].

*Corresponding Author: Ling Xia Liao; Email: liaolx@guat.edu.cn
DOI: https://doi.org/10.70003/160792642026052703003

Table 1. The related surveys

Surveys	Domains	Contributions
[1, 14]	General purpose	Review pre-trained LLMs and their major applications
[2]	Biomedical	summarize the progress of LLMs and applications in the biomedical domain
[5]	Education	present the potential benefits and challenges of educational applications based on LLMs from student and teacher perspectives
[4]	Business	Review the potential applications and issues of ChatGPT in supply chains
[15]	Software Engineering	Present the design techniques in the form of prompt patterns, solve the problems when using LLMs to automate common software engineering activities
[3]	Writing	Evaluate 15 abstracts generated by ChatGPT based on their titles and journals
[11]	Law	Survey of methodologies and application of infusion of legal knowledge to LLMs
[12]		Exploration of prompt engineering methodologies to increase the reasoning capabilities of legal LLMs
[13]		Evolvement of Legal LLMs and applications
[10, 16]		Surveys of legal LLMs and applications based on pre-defined key research questions

Challenges. Current literature also has not reviewed challenges in deployment. Legal texts are not only unusually long but also highly structured, interlinked, and open to interpretion. Contextual reasoning is problematic across jurisdictions and competing interpretations. High-stake risks arise from hallucination, from invented precedent to violations of due process. Since much of legal data remains proprietary or refstricted by copyright, data access and quality can be even more problematic than those in biomedical and financial domains.

Deploying legal LLMs has direct implications for justice, fairness, and democratic governance. Models trained on biased or incomplete legal data may perpetuate inequalities, and unverified outputs raise concerns about the unauthorized practice of law and professional liability. More important, literature reviews since Greco & Tagarelli [9] remain fragmented, with limited overlap across model, task, and dataset taxonomies.

Proposed Taxonomy. For legal LLMs, to provide a clear understanding of model structure in task performance and challenges for future directions, this paper selected 80+ publications by times cited, publication month (November 2022 to August 2025), journal, and dataset attributes. By systematically evaluating instances of domain-specific structure and dataset, this paper presents a comprehensive survey of legal LLMs in domain- and task-specific applications. The goal is to categorize model structure and applications in plain language, highlighting performance improvements, ethical risks, and practitioner needs to guide research collaboration and help streamline legal workflows toward client-centered service and judicial economy.

Contribution. The major contributions of this survey are as follows.

1. Model class: This survey synthesizes and simplifies group categories into three model classes: general structure, domain-specific, and commercial. Common model structures are presented and analyzed for each class to help NLP and legal researchers understand basic LLM concepts and choice of LLM for their purpose [9-10].
2. Legal task class: Simplifying Siino et al [16] and Sheik et al [12], this survey bases its four domain-specific classes on stages in legal workflow: access, structure, reason, and decision. This survey puts Legal Information Retrieval (LIR) and Legal Text Similarity (LTS) under access. Structure includes Named Entity Recognition (NER) and Legal Text Classification (LTC). Reason includes Legal Text Entailment (LTE), Legal Text Summarization (LTSU), and Legal Question Answering (LQA). The decision category includes case outcome prediction (COP) and draft generation (DG). This categorization helps AI researchers understand how Transformer-based LLMs fit into typical legal scenarios. [10, 13].
3. Dataset attribute: This survey selected 25 datasets. Besides jurisdiction, other attributes include times cited, areas of law, and data size, and open-access or closed.
4. We propose open challenges, research directions, and areas of future research related to exploring legal LLMs and their applications.

The rest of this survey is organized as follows: Section 2 presents related surveys. Section 3 introduces the model classes involving group categories and common models. Section 4 presents the legal task classes and domain-specific models and performance. Section 5 provides legal datasets. Section 6 discusses open challenges, research directions, and areas of future research, and Section 7 concludes the survey.

2 Related Surveys

The use of Transformer-based LLMs has long been a hot topic, given Table 1. After Raiaan et al. [1] and Wang et al. [14] who reviewed model and application and their general purpose, Wang et al. [2] reviewed biomedicine and their general pre-trained LLMs; Kasneci et al. [5] focused on education and the GPT-based; Frederic fato [4] summarized those for business management; White et al. [15] reviewed software engineering; and Gao et al. [3] directed their attention to academic & science writing. For the legal domain, the review upsurge was 2023, with Padiu et al.'s [10] five questions (i.e. leading LLMs to dataset) and with

Greco & Tagarelli's [9] multi-level group categories from model architecture to application, i.e. legal task class.

But subsequent authors did not follow this group categorization. Lai et al. (2024) [13] limited model architecture to the fine-tuned approach and the recently popular and application to tasks in judicial AI. A year later, Liu et al. [11] simplified categories even more and focused on approaches to knowledge infusion. And Sheik et al. [12] proceeded to ungroup Greco & Tagarelli on task class for model architectures involving prompt-based LLMs. The closest was Siino et al [16] (Dec 2024), who followed Greco & Tagarelli [9] on task classification and reviewed 61 publications, including about datasets. Following that task categorization (2023), this survey simplifies Greco & Tagarelli [9] on model architecture and application; simplifies Siino et al [16] with stages of legal workflow to group tasks, and categorizes dataset attributes in a separate section.

3 Transformer Architecture and Transformer-based Legal LLMs

In the legal domain, LLMs are often based on Transformer models. To better understand the typologies of Transformer model classes, training strategies, and user scenarios, this survey classifies legal LLMs into general-purpose models, domain-specific application, and commercial legal LLM application.

3.1 Transformer Architecture and Basic Models

Transformer models are the standard de-facto for Natural Language Processing (NLP) tasks. Transformer's core is its self-attention mechanism that uses parallel computation to optimize training efficiency and long-text dependency [17]. Transformer architecture consists of five core building blocks: (1) **Input representation** first converts each work/token into an embedding vector, as word positions encoded via positional encodings. (2) **Self-attention mechanism** then pinpoints the most relevant words after computing attention weights. Calculating attention multiple times in parallel, (3) **multi-head attention** would capture relationships under multiple heads such as syntax and semantics. (4) **Feed-forward network** then expands dimensionality before projecting back down, using two-layered MLPs applied separately to each token. Lastly, (5) **residual connections** prevent vanishing gradients and training is stabilized by layer normalization.

Original Transformer has an encoder-decoder structure. The encoder is a stack of layers with self-attentions and feed forward networks; the decoder is a similar with two kinds of self-attention – one masked to prevent looking ahead in generation and the other is cross-attention that attends to encoder outputs for translation. As shown in Figure 1 (Left), the encoder component incorporates multi-head self-attention mechanisms; the decoder component (right) integrates two attention mechanisms, one is masked multi-head self-attention for decoder outputs and the other is multi-head attention for encoder–decoder outputs.

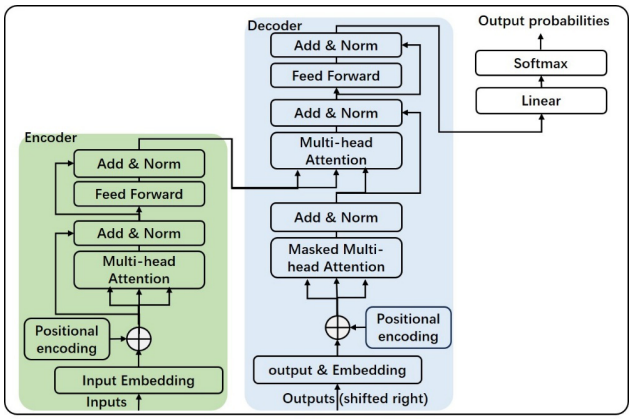


Figure 1. The original Transformer architecture

3.2 General-Purpose Transformer Models

This section categorizes general Transformer models into three classes: Decoder-only models, Encoder-only models, and Encoder-Decoder models. In Figure 2, this section shows each Transformer model class with common state-of-the-art examples. And as illustrated in Figures 3 to 6, this section displays model classes in the order of their respective architecture: Decoder, Encoder, Encoder-Decoder.

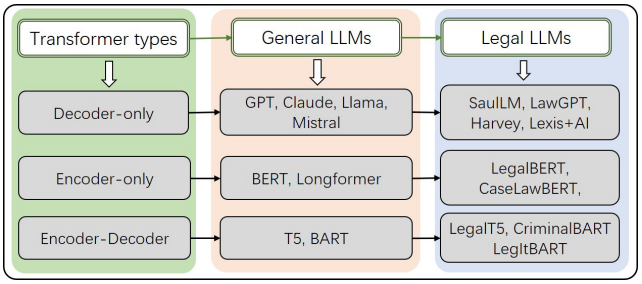


Figure 2. Category of Transformers in law & legal domain

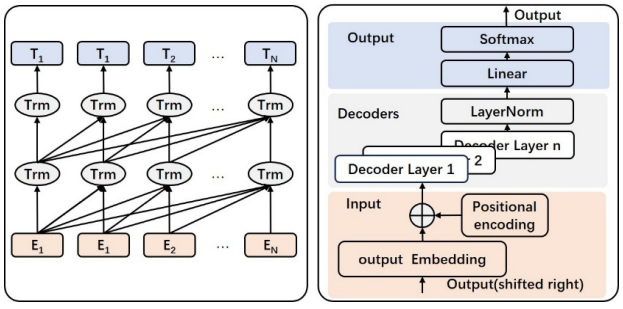


Figure 3. GPT architecture

With no encoder component in the original Transformer, the Decoder-only family models predict next tokens given previous ones, allowing tokens to attend only to themselves and preceding tokens. As shown in Figure 3, family differentiation is in the block feed-forward and connect-normalize layers. Feed-forward layers allow each block process features through full position-wise connection following self-attention. Next, layer normalization and residual connections stabilize training and improve gradient flow at multiple points. Such models

show high performance in text generation and completion, in legal tasks such as document classification, retrieval, and reasoning, whether used with prompt or adapter, or fine-tuned.

Encoder-based family models such as BERT [18] show results in sentence- and word-level tasks and can be adapted to downstream legal applications, such as legal question answering and named entity recognition, under bidirectional encoder Transformer architecture, as shown in Figure 4. BERT captures global semantic information by generating word vectors that jointly model both forward and backward contexts during pre-training.

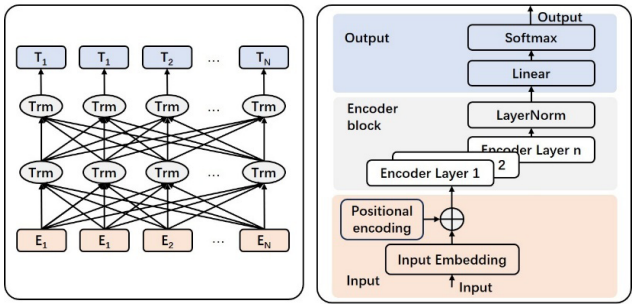


Figure 4. BERT architecture

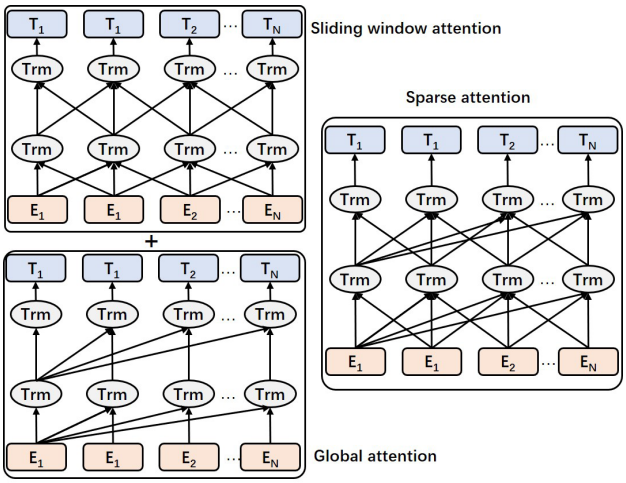


Figure 5. Longformer's sparse attention mechanism

Longformer [19] is an Encoder-based Transformer model that outperforms standard Transformers in long document sequences, by reducing the quadratic cost of self-attention. Because standard Transformer self-attention has complexity O(n^2) (encoder block: n unit length of sequence), Longformer introduces efficient sparse attention mechanism, as shown in Figure 5, and achieves complexity of O(n) or O(nlogn) for very long inputs. This sparse attention pattern involves sliding window attention and global attention. Local attention captures local dependencies at linear cost O(nw), where w = window size by allowing each token to attend only to a fixed window of nearby tokens. Global attention provides the model with global context, enabling important tokens to attend to all tokens, and allowing all tokens to attend back if needed. Longformer is now a common backbone for task-specific application.

Encoder-Decoder Transformer family models have an encoder stack and a decoder stack, as shown in Figure 6. The Encoder stack outputs contextual embeddings from source sequence inputs; the decoder stack predicts the distribution of next tokens from target sequence inputs. Using bidirectional encoder attention, causal decoder attention, cross-attention bridge, and positional encoding, Encoder-Decoder models show strong results in translation, summarization, text-to-text transfer, question answering, and code generation.

T5 and BART are common encoder-decoder models. T5 [20] reformulates all NLP tasks as text-to-text and shows strong performance in transfer learning tasks including translation, summarization, QA, and classification. BART [21] uses arbitrary type of document corruption including change in length, by designing a denoising autoencoder for pretraining. BART is very strong in abstractive summarization, robust to noisy input, and effective in sequence generation.

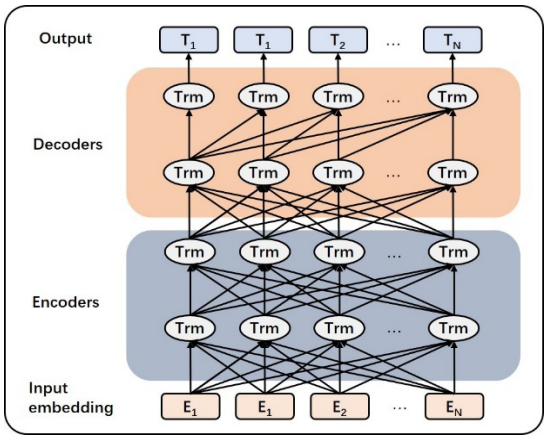


Figure 6. Encoder-Decoder model architecture

3.3 Domain-specific Application

Such legal LLMs are pre-trained for long legal documents in subdomains of law. Current models are often fine-tuned from BERT and Longformer architectures. Because broader legal context is more about legal order than model type, this section categorizes domain by continent of data origin: Multi-Continent, North America, European Union, and East Asia.

3.3.1 Multi-Continent Dataset

SaulLM [8] series is a Mistral-based LLM family designed by AI company Equall for legal reading comprehension and text generation. Pre-trained on a 500B-token base corpus, this model outperformed similarly sized models in summarization on documents from English databases from EDGAR to EuroParl.

3.3.2 North American Dataset

CriminalBART [22] is a BART-based Criminal Law Model for generating headnotes and summarizing judgments, with results in semantic context identification for French-Canadian criminal judgments. BERT-variant models include Custom Legal-BERT on US court decisions (1965-2020), and LegalPro-BERT on US SEC material contracts (2016-2019). LegalPro-BERT outperformed the BERT-Base in in-domain classification, using a

hyperparameter search space.

3.3.3 European Union Dataset

For token lengths exceeding the maximum, LegItBART [23] is a BART-based model for abstractive summarization on Italian legislation. The model showed improvement in summarization with a global-sparse-local attention mechanism.

Built on top of Google's T5, LegalT5 [7] is an encoder-decoder model for risk identification across 25 types of US SEC contracts. Given baselines involving BART and Pegasus, the model showed higher semantic similarity and average classification scores when evaluated using K-Means clustering.

3.3.4 East Asian Dataset

LegalBERT [6] is a BERT-based model pre-trained on a custom vocabulary involving Indian Central Government legislation, and Supreme Court and High Court decisions (1950 to 2019). With significant pre-training, the model outperformed baselines in statute retrieval, semantic segmentation, and judgment prediction.

LawGPT [24] is the first open-source Chinese Legal Knowledge–Enhanced LLM (2025) for Chinese judgment prediction on question-answer pairs and court decisions in criminal and various other areas of law. Allowable for private access, the model outperformed LLaMA 7B in six out of eight tasks and was fine-tuned with the LoRA technique.

3.4 Commercial Legal LLM Applications

Commercial legal AI platforms built on LLMs are deployed products in law firms, courts, and legal-tech vendors. They typically combine LLMs (often fine-tuned or RAG-augmented) with legal databases, workflow automation, and compliance. They currently include Harvey, Vincent AI, and AlphaGPT.

Harvey (www.harvey.ai) is a leading legal-gen AI platform offering law-firm–customized LLM tools for contract drafting, due diligence, litigation prep, and Q&A. The platform excels in enterprise integration and compliance, first built on GPT-4 and now with RAG over proprietary legal databases.

Vincent AI (vlex.com/vincent-ai) is an AI legal assistant that integrated its xLex global database for 20+ workflows including multi-jurisdictional research, strategic response in litigation, and judge profiling. Integrating agnostic RAG and leading LLMs like Anthropic, the platform used verified citations, and zero-data retention.

AlphaGPT (https://www.icourt.cc/product/alphaGPT) is a Chinese AI platform that integrated iCourt's Alpha with DeepSeek R1 for agentic workflows from proactive relationship development to client action timelines. Built with advanced legal knowledge graphs, the platform excels in legal information entailment and comparative data visualization.

4 Task-Specific Applications

To help with model choice in experimental design or in-house workflow, this section departs from Siino et al [16] and categorizes legal LLM tasks into Access, Structure, Reason, and Decision, mirroring practitioner workflows. For each scenario, a brief explanation is provided and the common LLM-based applications are summarized as shown in Figure 7.

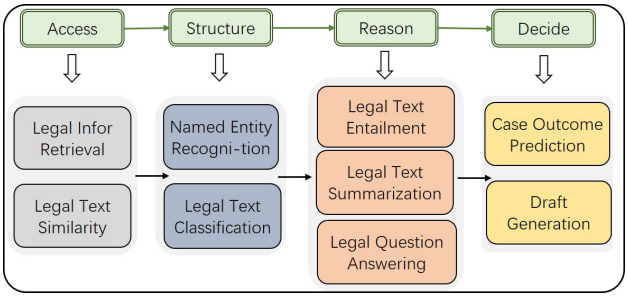


Figure 7. Legal work flow and the relevant application involving LLMs

4.1 Access

Access is the workflow stage of identifying and retrieving relevant statutes, court decisions, and practice guides. For AI/NLP, this process includes the applications of LIR and LTS, in Table 2.

4.1.1 LIR

In LIR, the goal is to have LLMs retrieve court decisions that follow or statute sections that inform the decision or section queried [9]. Despite experience in database searches, senior researchers take hours in engineering queries due to poor recall from missing terms, relevance ranking, and difficulties in identifying cross-jurisdictional common threads. On the other hand, legal LLMs excel in query suggestion, summarization, and other relevant similarities for retrieval. More specific, LIR often involves processing and understanding long legal documents. Common metrics include NDCG@K. Common applications are provided below by continent of data origin.

East Asia. Li et al. [25] proposed SAILER for the comprehension challenge with long sequence dependencies on Chinese criminal case datasets such as CAIL2022-LCR. SAILER used structural information and an asymmetric encoder-decoder architecture. SAILER used other metrics like MRR and outperformed various Chinese BERT baselines and Longformer in discriminative ability. Lee and Lee [26] proposed the Taiwan Longformer for long distance dependencies and evaluation on Taiwanese civil and criminal court decisions. The model embedded decision contexts with Longformer sliding window attention, and captured dependencies with LSTM gating mechanism. The retrieval model was evaluated on the AERC metric and surpassed CKIP BERT-BASE-Chinese and Longformer.

European Union. Colombo et al. [27] proposed the ETL pipeline for unstructured representation for Italian statutes. ETL enriched BERT and Mistral extraction by incorporating a statute property graph schema, LLMs, and a quality assessment of the resulting knowledge graph.

Multi-Continent. Nguyen et al. [28] took on the concision challenge utilizing small models and LLMs

and proposed a three-stage Retrieve–Revise–Refine framework, for the COLIEE 2022 (Japanese civil code bar exam) and 2023 (Federal Court of Canada) datasets. The framework significantly enhanced retrieval over state-of-the-art methods. The framework concisely identifies statute sections for entailment with queries.

Particular Jurisdiction. Phung et al. [29] addressed the language complexities of legal document retrieval by proposing a fine-tuning framework with synthetic training data generated by LLMs. The framework achieved advanced level of comprehension and data availability modes, by embedding relationships between questions, context, and answers.

Eggmann et al. [30] targeted legal citation recognition in large-scale document processing in a resource constrained environment and proposed Transformer-based encoder only models. The proposed models can process 3,000 documents in about 32 minutes, exceeding the performance requirements and ensuring scalability and efficiency of large scale legal platforms.

Althammer et al. [31] described the challenges of legal case retrieval including the domain specific language, the notion of relevance on paragraph-level, and the long documents. Two different approaches for handling longer documents with BERT were proposed. The paragraph level retrieval approach can achieve a high recall in the first stage retrieval, and the summarization approach can re-rank the retrieved results using BERT.

Table 2. The summary of the legal applications in the Access stage

App.	Methods	LLMs	Datasets	Descriptions
LIR	[26]	Longformer-LSTM	Taiwan legal cases	effectively understand and process long legal case documents
	[27]	BERT & Mistral-7B	a dataset of 45k laws	modeling the Italian legislation in a KG by employing fine-tuned LLMs, enrich and augment the KG, allow in-depth analysis of legislative data.
	[30]	Bert-based	MultiLegalPile dataset	recognize citations, ensure scalability and efficiency of largescale legal platforms
	[31]	Bert-based	COLIEE 2021	Bert-based LLMs for long legal documents,
	[25]	SAILER	LeCaRD CAIL2022-LCR COLIEE2020/2021	SAILER significantly outperform previous state-of-the-art methods in Chinese legal case retrieval.
	[29]	ChatGPT	Vietnamese legal datasets	generate synthetic training data, offer a scalable and effective solution to the limitations of existing pre-trained embedding models
	[32]	Transformer-based	COLIEE 2022	obtain knowledge from textbooks and web resources about the situational use of statutes
	[28]		COLIEE 2022/2023	pinpoint the concise set of legal articles that either entail a query or its negation.
LTS	[34]	16 models	Brazilian Portuguese legal dataset	evaluate 16 methods for calculating similarity in Brazilian Portuguese legal data
	[33]	Bert-based models	U.S. Patent Phrase to Phrase Matching dataset	Propose an ensemble approach incorporating 4 BERT-related models for patent document processing

Wehnert et al. [32] first imitated the process of a legal expert studying the situational application of statutes to infer relevance and entailment relationships between a query statement and a statute, then use Transformer-based architectures to extract additional statute information from textbooks and incorporate this knowledge into the original pipeline. The results indicated that there is a benefit of using textbook knowledge in Statute Retrieval and Entailment Classification tasks.

4.1.2 LTS

LTS is the process of determining how closely related or alike are two pieces of legal text. Problems with traditional LTS methods involve lexical overlap and missed legal semantics, with word embeddings such as Word2Vec and Glove. LLMs-based methods can use Transformer embedding and can better capture deeper legal meaning in synonym, paraphrase, and legalese. Common applications are as follows. A standard metric is cosine similarity.

North America. For cooperative Patent classification problems across language barriers and patent specification relationships, Yu et al. [33] proposed an ensemble approach on the US Patent Matching dataset. The BERT-based models and ELECTRA achieved a 0.8534 CV score on the Pearson correlation coefficient. In South America, Silva et al. [34] focused on the same query anchor document and evaluated 16 methods on Brazilian Portuguese legal documents. Also on Jensen-shannon divergence, results show that text representation return a different set of documents.

4.2 Structure

Structure is the workflow stage of drawing connections among pieces of legal information, such as legal entities, legal topics, and sentencing disparities. Corresponding AI/NLP tasks include NER and LTC, applications shown in Table 3.

4.2.1 NER

NER identifies and classifies textual entities into predefined categories. With longer-range dependencies, legal NER targets domain-specific entities (legal citation, presiding judge name, or legal principle) that are nested and irregular. More than rule-based, LLM adapt on a label-basis. Common metrics include **precision (P), recall (R), and F1 score**. Common applications below are outlined by continent of data origin.

Multi-Continent. Deuber et al. [35] evaluated 11 legal LLMs on seven NER datasets and also used micro-F1. Results showed that GPT-4o outperformed others in few-shot settings across all dataset languages except German, and on datasets involving U.S. SEC filings and U.S. court decisionsThe authors provided valuable benchmarks for legal NER applications.

Southeast Asia. Given open-source metadata on Indian court decisions, Kalamkar et al. [36] introduced a new corpus and BERT-based baselines, to extract 46545 annotated named entities mapped to 14 types. Among baseline models, the RoBERTA model with transition-based parser achieved the highest score in recall, precision, and F1, using a Spacy pipeline.

East Asia. Zhang et al. [37] proposed a RoBERTa-Global Pointer-based method to extract Chinese feature representations on the CAIL 2021 dataset (burglary court decisions). The model outperformed ROBERTa models and Lawformer, using three aspects. The model captures contextual entities with RoBERTa for the char-level, the Skip-Gram method for the word-level, and the GlobalPointer method with softmax function to weigh entity type scores.

Table 3. Summary of legal applications in the Structure stage

App.	Methods	LLMs	Datasets	Descriptions
NER	[36]	BERT	Indian legal named entities	introduce a new corpus of 46545 annotated Indian legal named entities mapped to 14 legal entity types with baseline models
	[37]	RoBERTa	Chinese judicial domain dataset	proposed a RoBERTa-GlobalPointer-based method for NER of legal documents
	[35]	11 LLMs	7 legal datasets in diff. languages	conduct a comparative analysis of 11 state-of-the-art LLMs on legal NER tasks across 7 diverse datasets
	[38]	Non-LLMs	Uzbek legal dataset	present a dataset and approaches to named entity recognition in Uzbek language
	[39]	ELMo BERT	HAREM LeNER-Br DrugSeizures-Br	explore different scenarios considering two deep language architectures (ELMo and BERT), four unlabeled corpora and three legal NER tasks for the Portuguese language
LTC	[41]	GPT-3	LegalDocML LegalRuleML	Fine-tuned GPT-3 can recognize the difference between obligation, permission, and constitutive rules with performances overcoming previous scores in legal rule classification
	[43]	GPT- Bert-based	KICS	models struggle to interpret implicit contextual cues within legal texts. High performance and interpretability in legal AI models are needed.
	[42]	BiLSTM	online public legal database	lengthy legal documents classification
	[40]	Transformer-based	JRC-Acquis EURLEX57K l	study the performance of various recent Transformer-based models in combination with strategies for large multi-label text classification

Particular Jurisdiction. Mengliev et al. [38] presented two algorithms and showed results over Uzbek language dataset, including 1,160 sentences with nearly 19,000 word forms. Bonifacio et al. [39] fine-tuned LLMs on Portuguese large intradomain corpora and found that they showed significant improvement in performance. With two deep language architectures (ELMo and BERT), the authors used four unlabeled corpora and three legal NER tasks for the Portuguese language.

4.2.2 LTC

LTC classifies legal texts into predefined categories, including document type, legal topic, court procedure, case outcome, and rhetorical roles. More than TF–IDF or Naïve Bayes, current LTC adopts LLM- and GPT-based classifiers such as CaseLawBERT. Common applications are listed as follows by continent of data origin. The standards criteria are classification metrics.

European Union. To address problems in dependencies in multi-label classification, Shaheen et al. [40] constructed datasets JRC-Acquis and EURLEX57K on EU legislation labeled with ~7000 EuroVoc concepts, and evasluated BERT-based models and XLNet. The authors chose other metrics like NDCG@K and variants like micro-F1 and RP@K. Results showed that BERT-based models outperformed the LSTM baseline with state-of-the-art results, using classification strategies such as generative pretraining, gradual unfreezing and discriminative learning rates**.**

Given the specific problems in the legal_rule classification task, Liga et al. [41] fine-tuned GPT-3 to identify rule distinctions from 707 provisions from the European GDPR (General Data Protection Regulation). To tell rules apart from obligatory to constitutive, the authors used two XML standards to annotate legal documents and the metrics accuracy and F1.

Given other problems in classification of lengthy legal documents, Wan et al. [42] showed improved F1 performance with BiLSTM on US SEC filings. The authors' method involves dividing text into segments and combining the resulting embeddings to form a single document embedding.

East Asia. Lee et al. [43] conducted a comprehensive evaluation of various legal task models based on Korean sexual offense court decisions. The experimental results demonstrate that methods fine-tuning small language models such as KLUE-BERTon legal data yield superior performance compared to benchmark models large general-purpose models such as GPT-3.5 and GPT-4.0, as well as benchmark models traditional machine learning models.

Table 4. Summary of legal applications in the Reason stage

App	Methods	LLMs	Datasets	Descriptions
LTE	[47]	LLMs		analyze the causal relation structures, generate high-quality synthetic data, strengthen reasoning ability
	[46]	Bert-based	COLIEE 2022	ensemble a rule-based method and a BERT-based method in COLIEE 2022 task 4.
	[44]	BM25, Transformer	COLIEE 2021	Transformer-based approach for LTE, BM25 for LIR, semantic knowledge for statute law entailment task.
	[45]	BERT-based	COLIEE 2021	provide a data-augmentation method to make training data using civil code articles
LTS	[49]	Legal-LED Legal-Pegasus	UK & Indian Supreme Court judgement	Examine the performance of models and methods for legal case judgement summarization.
	[50]	Ontological model	administrative regulations	Form an ontological model to describe, organize and present the relationships between the facts and the infor extracted
	[48]	Mluti LLMs	CLSum	Propose CLSum for summarizing multi-jurisdictional common law court judgment documents
	[51]	LeSum	Indian Supreme Court & Calcutta High Court judgments	Propose a cost-effective hybrid summarization methodology for Indian legal judgments
	[52]	Transformer	Brazil's administrative court dataset	investigate the effectiveness of LLMs in conjunction with state-of-the-art embeddings for legal precedents retrieving
LQA	[53]	LLaMA- 2-13B	Syn- LeQA	Propose AceAttorney, an LLM distillation approach, to develop LQA data and supervised models without human annotation.
	[56]	Llama-2-7b-hf	IndicLegalQA	presents a comprehensive dataset for LQA in the Indian judiciary context that facilitates
	[54]	DiscoLQA	Q4EU	investigate mechanisms to capture the patterns of legalese for zero-shot question answering
	[55]	UniLM, T5, BART	cLegal-QA,	provide accurate and timely answers to Chinese legal questions

4.3 Reason

Reason in legal workflow is the stage of interpreting and applying legal information to analogize and distinguish court decisions. As shown in Table 4, The corresponding AI/NLP tasks include LTE, LTS, and LQA.

4.3.1 LTE

LTE determines whether a legal statement or holding is implied from other legal arguments, whether formulated from statute, doctrines, or procedures. Problems in entailment are ambiguity, jurisdictional variance, conflicting precedents, and context sensitivity. If fine-tuned, LLMs can classify legal premises from hypothesis and validate arguments, to help in summarization and extraction. Common metrics include accuracy and F1. Common application are provided below.

East Asia. For the 2021 COLIEE in statute entailment task, Kim et al. [44] applied a TDIF model on Japanese Civil Code articles. The model used a document length control mechanism and applied semantic knowledge to statute sections. The model won 3rd place on task 2. For the same task in COLIEE 2021, Aoki et al. [45] proposed a BERT-based ensemble model with 0.7037 accuracy, outperforming 10 BERT test models and seven other submitted models. The models enriched syntactic structure of the questions and articles by training data augmentation. Similarly as Task 4 in COLIEE 2022, Fujita et al. [46] proposed the Rule- and BERT-based method and achieved correct answer ratio 0.6789, best among submissions. Chu et al. [47] leveraged LLMs to improve reasoning capabilities for entailment. The models outperform previous methods on high-quality synthetic data. The authors used generated synthetic datasets to analyze the NLP task causal relation structures within legal documents.

4.3.2 LTS

LTS is the task of automatically generating brief legal text summaries, for example, involving key facts, issues, reasoning, and decisions. More than compressing information into new sentences, LTS outputs must be structured, neutral, precise, and traceable. Traditional methods show errors in length, faithfulness, granularity and precision. LLMs now enable structured, high-quality, plain-language summaries that align with legal workflows. Common applications are listed as follows by continent of data origin.

Multi-Continent. Liu et al. [48] constructed the first summarization dataset across common law jurisdictions (CLSum). The authors also proposed a legal

knowledge enhanced evaluation metric and baseline model. Results verified that the baseline can perform well in the few-shot and zero-shot settings, using data augmentation and summary generation. Deroy et al. [49] developed abstractive and extractive summarization models and applied various domain- and task-specific variants **to** court decisions of the UK and Indian Supreme Courts. Results show that abstractive summarization models and LLMs generally perform better. Limitations are inconsistencies and hallucinations in generative outputs.

European Union. Benedetto et al. [50] proposed LLMs to summarize Italian legal news documents. The models outperformed baselines such as BART and T5 in grammaticality and informativeness in a zero-shot setting.

Southeast Asia. Ansar et al. [51] introduced the LeSum hybrid model on Indian legal judgments. The model reduced 59% token count for extractive summarization and costs, and tuning costs over four baseline LLMs: PEGASUS, BART, Gemini and Llama. For extractive summarization, the model used a novel topic-supervised graph-based context-matching mechanism to capture key information with minimal redundancy. And for abstractive summarization, the model used a zero-shot approach to generate structurally-segmented summaries without requiring annotated samples. Mentzingen et al. [52] studied the interplay between NLP methods, text summarization techniques, and embeddings from language models in constructing expert systems dedicated to the legal task retrieval of legal precedents, with an emphasis on achieving cost-efficiency problem. They used methods various summarization methods to distill legal document corpora of different types into a convenient form that retains their essence and investigated the effectiveness of these methods in conjunction with state-of-the-art embeddings performance benchmark models based on LLMs. The results suggest that while the full text embedded with ADA's extensive context window leads in retrieval performance, a balanced combination of metric POS-derived summaries and ADA embeddings presents a compelling trade-off between retrieval performance and resource cost.

4.3.3 LQA

LQA is the task of taking natural language queries and answering with legal sources such as cases, statutes, regulations, and contracts. This section categorizes LQA in the reason stage because LQA requires more than retrieval; LQA reasons whether the retrieved text is an accurate answer. LQA methods range from rule-based expert systems to information retrieval systems that fetch relevant cases/statutes, to machine reading comprehension, and to LLMs fine-tuned on legal corpora. Common question answer applications based on LLMs are as below by continent of data origin. A standard metric is MRR.

North America. To develop supervised LQA models without manual annotation, Italiani et al. [53] presented AceAttorney and the Syn-LeQA, a synthetic dataset of US civil rights court decisions, Google Store privacy policies, and site terms of service. Results showed effectiveness and efficiency over baseline models, using a selective generative paradigm. The paradigm retrieved documents given retrieval-based scenarios.

European Union. Sovrano et al. [54] modeled for discursive patterns in legalese and proposed a model for zero-shot question answering. Evaluated on other metrics such as NDCG, P, and F1, the model outperformed the baselines with two aspects. The model first optimized retrieval for change in information type, pre-trained on open-domain answer retrieval systems, with **a discourse-aware selection method**

East Asia. Wang et al. [55] developed the first large-scale dataset for Chinese legal question answering and model, on cLegal-QA, 14,000 high-frequency questions about civil law. Results on variants of ROUGE and BLEU showed 60–80% accuracy over the baseline model.

Southeast Asia. For lack of an Indian LQA dataset, Veningston et al. [56] present an approach and the IndicLegalQA dataset, court decisions in criminal and civil law.

4.4 Decision

Decision is the workflow stage of predicting legal decisions, in tasks from client consultation to case merit assessment. Given research attention, corresponding AI/NLP applications consist of COP and DG, listed in Table 5.

4.4.1 COP

COP is the task of forecasting court decisions, often used in case-merit assessment. More task-specific than traditional deep-learning networks, LLM-based models can be prompt-based or fine-tuned for court order and risk labels based on legal reasoning. Common COP applications are the following by continent of data origin. Standard evaluation metrics are AUC or standard classification metrics.

Multi-Continent. To evaluate structured-reasoning accuracy on domain-specific datasets,Wang et al. [57] introduced the LegalReasoner framework, on the Chinese AI and Law Challenge (CAIL2018) dataset and ECHR court decisions. The framework demonstrated substantial improvements over state-of-the-art baselines in four stages. The framework started with legal knowledge infusion and pre-trained LLMs with contrastive learning techniques. Next, the framework used GNN to retrieve relevant statutes and court decisions. Third with Transformer-based architecture, the framework used multi-hop reasoning on balancing or bright-line tests, using a hierarchical attention mechanism. Lastly, the framework generated arguments by synthesis, using a generative adversarial network.

Le et al. [58] adopted Transformer-based pre-trained models to capture fine-grained semantic interaction among complicated legal elements. They collocated distinct special tokens for each legal element, and extracted the case features from different perspectives. To utilize the structural information of the case, numerous experiments are carried out on a public real-world dataset. The proposed approach achieved competitive performance on standard task metrics and MAE, compared to the currently published state-of-the-art method.

Wei et al. [59] proposed a LLM-based neuro-symbolic framework for legal task legal judgment prediction. They also developed a Chain-of-Thought prompt that uses

LLMs to extract fact elements from a legal-case corpus. These elements act as logical variables within the rules, supporting the reasoning process in the logic module and improving overall interpretability metrics.

European Union. Trautmann et al. [60] trained a prediction model over deicions of ECHR and the Federal Supreme Court of Switzerland. The model achieved better classification results over the baseline, using Legal Prompt Engineering. Masala et al. [61] focused on the legal domain and introduced a Romanian BERT model pre-trained on a large specialized corpus. The proposed model outperforms several strong prediction baselines on AUC and corpora consisting of bank trial cases from Romania.

East Asia. Sun et al. [62] focused on PLM prompt templates and proposed Lawformer on Chinese legal documents. The model outperformed the baselines in low-resource and data-rich scenarios, using a knowledge base. The model incorporated key information in soft prompt tokens, which are allowed to exceed 512 tokens. Su et al. [63] aimed to predict judgment from case facts and introduced the JuDGE (Judgment Document Generation Evaluation) benchmark, on Chinese legal documents. Results on other metrics also showed strong performance over baselines also on BERTSCORE and METEROR, by pairing case facts with their corresponding full-length judgments.

Li et al. [64] proposed a BERT-based approach for predicting court decision text-to-text from case facts, on two Chinese datasets. Also on BLEU, the model showed higher performance over baselines, by using two separate optimizers in training. Integrating sentence and statute section, the model used BERT encoder and Transformer decoder.

Table 5. Summary of legal application in the Decision stage

App	Method	LLMs	Datasets	Descriptions
COP	[59]	GPT-4	a large dataset of private lending cases	incorporate symbolic reasoning into the neural networks and propose a new neural-symbolic legal judgment prediction framework for civil case based on LLMs
	[60]	GPT	ECHR&Switzerland Federal Supreme Court cases	use LPE with LLMs over long legal documents for the Legal Judgement Prediction (LJP) task.
	[57]	LLaMA2	ECHR Cases CAIL2018	propose a novel multi-stage framework that integrates LLMs, domain-specific knowledge, and structured reasoning for enhanced legal judgment prediction
	[61]	Bert based	RoJur RoBanking	introduce a Romanian BERT model pre-trained on a large specialized corpus for legal judgement prediction
	[62]	Lawformer	CAIL2018	leverage a pre-trained Language Model for Chinese legal judgment prediction task
	[58]	Lawfor-mer	300 law articles	capture the fine-grained semantic interactions among different legal elements, share parameters among modules
GD	[65]	Multi-LLMs	CaseGen	introduce a benchmark for multi-stage legal case documents generation in the Chinese legal domain
	[66]	Non- LLMs	administrative regulations	Ontology based features for attributes extraction, semantic search to improve the functionality of NLP in legal domain
	[67]	Non- LLMs	10000 Chinese legal docs	propose a text style transfer model based on general adverse networks for legal documents generation
	[68]	Ransformer	docs from China Judgements Online	generates texts with slots and fills the slots using Transformer-based Key-Value Memory Networks.
	[63]	RAG	criminal case docs from China Judgments Online	introduce a novel benchmark for evaluating the performance of judgment document generation in the Chinese legal system
	[64]	Bert	Docs from China Judgements Online	automatically generate court view from the fact description of a legal case

4.4.2 DG

DG automates drafting of legal documents including contracts, pleadings, and compliance reports. This survey puts DG in the decision stage because the output is a version of a legal opinion memo beyond retrieval and classification. More than template-filling, LLM-based DG generates context-aware clauses such as confidentiality, arbitration, or termination with natural variation. The literature showed no standard metric.

Current DG models are often based on East Asian legal documents. Because of the lack of benchmark, Li et al. [65] introduced CaseGen, a benchmark for case document prediction (multi-stage), on annotated Chinese court decisions and common section labels. The benchmark is applicable to drafting tasks in defense statements, trial facts, legal reasoning, and judgment results. But the benchmark the authors proposed cannot be used for several widely used domain-specific and task-specific models.

Ivaschenko et al. [66] proposed a subject area ontological model, pre-trained on Chinese administrative regulations. The model outperformed the baselines on entity extraction, knowledge-based text understanding, and text generation. The model extracted text attributes from normative legal acts to named entities and their relationships.

Li et al. [67] proposed a text style transfer model tst-gan for ideal legal document prediction, on a Chinese dataset with reduced workload over the baselines. With confrontation simulation, the model trained with discriminator and reformated text with text style transfer.

For coherent legal document generation, Huang et al. [68] proposed a novel method CoLMQA. Results show document quality over the baseline, using pure language modeling and question answering. The model first modeled text to generate slots and filled them with Transformer-based Key-Value Memory Networks.

5 Legal Datasets

Domain-specific legal LLMs rely on domain-specific datasets. Unlike general NLP corpora, legal benchmark datasets are fewer, confined to certain jurisdictions, and often restricted by copyright or other license. Nevertheless, as the literature provides, this section will enumerate high-performance datasets based on the jurisdiction of data origin, applicable to stages from training to evaluation, from classification to prediction. Under data origin, this section further categorizes datasets under workflow stage and attribute.

North America. Most benchmark datasets are from the US. This section lists 11 datasets.

- Access: For the retrieval task, datasets involve MAUD (merger agreement) [16], CUAD [7] andContractNLI (contract) [9] , and USPTO-2M (patent) [69].
- Structure: For the NER task, sets involv LegalDB (times cited), and, under specific areas of law, ALeaseBERT and LexGlue.
- Reason: For entailment (areas of law), sets include LexFile (all), and BillSum (all) for summarization. And for LQA, sets involve CaseHOLD (times cited) and Syn-LeQA (specific areas of law).

European Union. Datasets are often multi-language for member states. This section lists four datasets.

- Access: For the retrieval task (area of law), the set is from Normattiva [27].
- Reason: For LQA (jurisdiction), the set is LEXTREME [70].
- Decision: For prediction (areas of law), the set is 2UK2EU (all) [71] and JRC-Acquis (all) [40].

East Asia. This section lists five Chinese and Indian legal sets.

- Access: For retrieval task (times cited), the set is LeCaRD [72].
- Structure: For the NER task, the set is Lawbench (areas of law) [73] and Indian Kanoon [56].
- Reason: for LQA (size), the set involves SLARD [74] and cLegal-QA [55].

Multi-Continent. this section lists two sets.

- Access and Structure: For classification and retrieval tasks (size), the set is COLIEE (Japanese and Canadian) [28].
- Access, Reason and Decision: For retrieval, LQA and prediction tasks (times cited), the set is Lawformer Chinese [62].

6 Challenges and Open Issues

Investigation of legal LLMs and their applications has been a hot topic of interest in the NLP and legal domains. However, this research has faced open challenges and research issues.

Regarding the domain-specific challenges, this research firstly faces the challenge of access to high-quality training dataset, because these datasets are often proprietary and government legal databases lack area-specific labels. Another challenge is longer-range context embedding, even given techniques such as input context length exceeding 1024, decoders with stronger discriminatory vectors, and adaptive RAG [25].

A third challenge is the choice of evaluation metric because no author has reviewed the data factors that determine the choice of metric variants, for which task is the standard, and even environmental impact, not to mention complications in subdomain-specific sets (CaseHOLD, CUAD, LexGLUE, LegalBench).

Legal LLMs have hallucination challenges in prediction. A related problem is post-processing. Models on unclear queries lead to incorrect legal interpretation, or fake court decisions and source citation. And engineering solutions have to fit into practitioner workflow.

For ethical and governance challenges, one is data representation in factors such as systemic race, economic status, and availability relating to litigant bias [13]. Another is how to define terms of professional conduct such as solicitor-client privilege, judicial apprehension of bias, and fraud in data transfer, especially for cases in private international law, even for firms that have developed their model in-house [13]. Another challenge is global LLM disclaimer and safeguards, for training and fine-tuning across jurisdictions.

Respective research issues involve long-context modeling and dataset construction; performance metrics; retrieval-augmented generation pipelines reducing hallucination; and pipelines in compliance with accountability obligations for explanability and data protection.

7 Conclusion

This survey reviewed Transformer-based Legal LLM categories, simplifying class groups for general model structure, specialized application, and relevant research issues. Task-specific models or datasets such as SAILER (Encoder-only), AceAttorney (Encoder-Decoder), and the CaseGen benchmark (Decoder-only) demonstrate strong performance in LIR, LQA, NER and COP. However, open challenges persist, including the limited availability of datasets (high quality or licensed) for subdomain

pre-training, the heterogeneous dependencies and key facts of legal documents, raw hallucination rates, and explainability and security safeguards. Current benchmarks insufficiently capture the context windows of legal writing logic, while ethical concerns raise questions about implementation principles, such as safety, bias, and context relevance. Addressing these issues requires advances in modeling context window length, sublanguage and macro-averaged measures, prompt engineering approach and RAG pipelines, and access to and amendment obligation for organizations with personal information, combined with close collaboration among NLP researchers and legal practitioners. Ultimately, Transformer-based Legal LLMs hold transformative potential for legal practice, but their accuracy, trustworthiness, and transparency must be safeguarded before adoption for client-centred service and judicial economy.

Acknowledgements

This work was supported by the the National Natural Science Foundation of China (61962016), Natural Science Foundation of Fujian (2023J011807, 2023J05309), Natural Science Foundation of Guangxi (2023GXNSFAA026018).

References

[1] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, S. Azam, A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges, *IEEE Access*, Vol. 12, pp. 26839-26874, February, 2024. https://doi.org/10.1109/ACCESS.2024.3365742

[2] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, J. Fu, Pre-trained language models in biomedical domain: A systematic survey, *ACM Computing Surveys*, Vol. 56, No. 3, pp. 1-52, March, 2024. https://doi.org/10.1145/3611651

[3] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, A. T. Pearson, Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers, *NPJ Digital Medicine*, Vol. 6, No. 1, Article No. 75, April, 2023. https://doi.org/10.1038/s41746-023-00819-6

[4] G. F. Frederico, ChatGPT in Supply Chains: Initial Evidence of Applications and Potential Research Agenda, *Logistics*, Vol. 7, No. 2, Article No. 26, April, 2023. https://doi.org/10.3390/logistics7020026

[5] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, *Learning and Individual Differences*, Vol. 103, Article No. 102274, April, 2023. https://doi.org/10.1016/j.lindif.2023.102274

[6] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The Muppets Straight Out of Law School, *arXiv preprint*, arXiv: 2010.02559, October, 2020. https://doi.org/10.48550/arXiv.2010.02559

[7] A. Poornima, K. V. Nagaraja, M. Venugopalan, Legal Contract Analysis and Risk Assessment Using Pre-Trained Legal-T5 and Law-GPT, *3rd International Conference on Integrated Circuits and Communication Systems*, Raichur, India, 2025, pp. 1-8. https://doi.org/10.1109/ICICACS65178.2025.10968817

[8] P. Colombo, T. Pires, M. Boudiaf, R. Melo, D. Culver, E. Malaboeuf, G. Hautreux, J. Charpentier, M. Desa, Saullm-54b & Saullm-141b: Scaling up Domain Adaptation for the Legal Domain, *NIPS'24: Proceedings of the 38th International Conference on Neural Information Processing System*, Vancouver BC Canada, 2024, pp. 129672-129695.

[9] C. M. Greco, A. Tagarelli, Bringing order into the realm of Transformer-based language models for artificial intelligence and law, *Artificial Intelligence and Law*, Vol. 32, No. 4, pp. 863-1010, December, 2024. https://doi.org/10.1007/s10506-023-09374-7

[10] B. Padiu, R. Iacob, T. Rebedea, M. Dascalu, To What Extent Have LLMs Reshaped the Legal Domain So Far? A Scoping Literature Review, *Information*, Vol. 15, No. 11, Article No. 662, November, 2024. https://doi.org/10.3390/info15110662

[11] Q. Liu, H. Yu, Q. Wang, Q. Xu, J. Li, Z. Zou, R. Mao, E. Cambria, Legal knowledge infusion for large language models: A survey, *Information Fusion*, Vol. 125, Article No. 103426, January, 2026. https://doi.org/10.1016/j.inffus.2025.103426

[12] R. Sheik, S. A. Reji, A. Sharon, M. A. Rai, S. J. Nirmala, Advancing prompt-based language models in the legal domain: adaptive strategies and research challenges, *Artificial Intelligence and Law*, pp. 1-43, May, 2025. https://doi.org/10.1007/s10506-025-09459-5

[13] J. Lai, W. Gan, J. Wu, Z. Qi, P. S. Yu, Large language models in law: A survey, *AI Open*, Vol. 5, pp. 181-196, 2024. https://doi.org/10.1016/j.aiopen.2024.09.002

[14] H. Wang, J. Li, H. Wu, E. Hovy, Y. Sun, Pre-trained language models and their applications, *Engineering*, Vol. 25, pp. 51-65, June, 2023. https://doi.org/10.1016/j.eng.2022.04.024

[15] A. Nguyen-Duc, P. Abrahamsson, F. Khomh, Eds., Generative AI for Effective Software Development, *Springer Nature Switzerland*, 2024.

[16] M. Siino, M. Falco, D. Croce, P. Rosso, Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches, *IEEE Access*, Vol. 13, pp. 18253-18276, January, 2025. https://doi.org/10.1109/ACCESS.2025.3533217

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach California USA, 2017, pp. 6000-6010.

[18] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA, 2019, pp. 4171-4186.

[19] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long-Document Transformer, *arXiv preprint*, arXiv: 2004.05150, December, 2020. https://doi.org/10.48550/arXiv.2004.05150

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, Vol. 21, No. 1, Article No. 140, January, 2020.

[21] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension, *arXiv preprint*, arXiv: 1910.13461, October, 2019. https://arxiv.org/abs/1910.13461

[22] N. Garneau, E. Gaumond, L. Lamontagne, P. Déziel, Criminelbart: A French Canadian Legal Language Model Specialized in Criminal Law, *18th International Conference on Artificial Intelligence and Law*, São Paulo, Brazil, 2021, pp. 256-257. https://doi.org/10.1145/3462757.3466147

[23] I. Benedetto, M. La Quatra, L. Cagliero, LegItBART: A Summarization Model for Italian Legal Documents, *Artificial Intelligence and Law*, pp. 1-31, February, 2025. https://doi.org/10.1007/s10506-025-09436-y

[24] Z. Zhou, J. X. Shi, P. X. Song, X. W. Yang, Y.X. Jin, L. Z. Guo, Y. F. Li, Lawgpt: A chinese legal knowledge-enhanced large language model, *arXiv preprint*, arXiv: 2406.04614, June, 2024. https://doi.org/10.48550/arXiv.2406.04614

[25] H. Li, Q. Ai, J. Chen, Q. Dong, Y. Wu, Y. Liu, C. Chen, Q. Tian, SAILER: Structure-Aware Pre-Trained Language Model for Legal Case Retrieval, *46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei, Taiwan, 2023, pp. 1035-1044. https://doi.org/10.1145/3539618.3591761

[26] H. Lee, H. Lee, Taiwan Legal Longformer: A Longformer-LSTM Model for Effective Legal Case Retrieval, *5th International Workshop on Artificial Intelligence and Education*, Tokyo, Japan, 2023, pp. 156-161. https://doi.org/10.1109/WAIE60568.2023.00036

[27] A. Colombo, A. Bernasconi, S. Ceri, An LLM-assisted ETL Pipeline to Build a High-quality Knowledge Graph of the Italian Legislation, *Information Processing & Management*, Vol. 62, No. 4, Article No. 104082, July, 2025. https://doi.org/10.1016/j.ipm.2025.104082

[28] C. Nguyen, P. Nguyen, L.-M. Nguyen, Retrieve–Revise–Refine: A Novel Framework for Retrieval of Concise Entailing Legal Article Set, *Information Processing & Management*, Vol. 62, No. 1, Article No. 103949, January, 2025. https://doi.org/10.1016/j.ipm.2024.103949

[29] N. H. Phung, C. T. Nguyen, M.-T. Nguyen, T. H. Nguyen, H. L. Le, T.-P. Nguyen, A Fine-tuning Framework Based on Question, Context, and Answer Relationships for Enhancing Legal Information Retrieval, *Engineering Applications of Artificial Intelligence*, Vol. 159, Part B, Article No. 111570, November, 2025. https://doi.org/10.1016/j.engappai.2025.111570

[30] M. Eggmann, J. Fitz, D. Glasl, M. Purandare, Citation Recognition in Large-Scale Legal Platforms Using Transformer Models, *2025 IEEE Swiss Conference on Data Science (SDS)*, Zürich, Switzerland, 2025, pp. 71-78. https://doi.org/10.1109/SDS66131.2025.00017

[31] S. Althammer, A. Askari, S. Verberne, A. Hanbury, DoSSIER@COLIEE 2021: Leveraging Dense Retrieval and Summarization-Based Re-Ranking for Case Law Retrieval, *arXiv preprint*, arXiv: 2108.03937, August, 2021. https://doi.org/10.48550/arXiv.2108.03937

[32] S. Wehnert, L. Kutty, E. W. De Luca, Using Textbook Knowledge for Statute Retrieval and Entailment Classification, *JSAI International Symposium on Artificial Intelligence*, Kyoto, Japan, 2022, pp. 125-137. https://doi.org/10.1007/978-3-031-29168-5_9

[33] L. Yu, B. Liu, Q. Lin, X. Zhao, C. Che, Semantic Similarity Matching for Patent Documents Using Ensemble BERT-Related Model and Novel Text Processing Method, *arXiv preprint*, arXiv: 2401.06782, January, 2024. https://doi.org/10.48550/arXiv.2401.06782

[34] D. Silva Junior, D. Oliveira, A. Paes, Evaluating Text Representations for Unsupervised Legal Semantic Textual Similarity in Brazilian Portuguese, *Discover Data*, Vol. 3, No. 1, Article No. 23, June, 2025. https://doi.org/10.1007/s44248-025-00052-4

[35] T. Deußer, C. Zhao, L. Sparrenberg, D. Uedelhoven, A. Berger, M. Pielka, L. Hillebrand, C. Bauckhage, R. Sifa, A Comparative Study of Large Language Models for Named Entity Recognition in the Legal Domain, *2024 IEEE International Conference on Big Data*, Washington, DC, USA, 2024, pp. 4737-4742. https://doi.org/10.1109/BigData62323.2024.10825695

[36] P. Kalamkar, A. Agarwal, A. Tiwari, S. Gupta, S. Karn, V. Raghavan, Named Entity Recognition in Indian Court Judgments, *arXiv preprint*, arXiv: 2211.03442, November, 2022. https://doi.org/10.48550/arXiv.2211.03442

[37] X. Zhang, X. Luo, J. Wu, A RoBERTa-GlobalPointer-Based Method for Named Entity Recognition of Legal Documents, *2023 International Joint Conference on Neural Networks*, Gold Coast, Australia, 2023, pp. 1-8. https://doi.org/10.1109/IJCNN54540.2023.10191275

[38] D. Mengliev, V. Barakhnin, N. Abdurakhmonova, M. Eshkulov, Developing Named Entity Recognition Algorithms for Uzbek: Dataset Insights and Implementation, *Data in Brief*, Vol. 54, Article No. 110413, June, 2024. https://doi.org/10.1016/j.dib.2024.110413

[39] L. H. Bonifacio, P. A. Vilela, G. R. Lobato, E. R. Fernandes, A Study on the Impact of Intradomain Finetuning of Deep Language Models for Legal Named Entity Recognition in Portuguese, *Brazilian Conference on Intelligent Systems*, Rio Grande, Brazil, 2020, pp. 648-662. https://doi.org/10.1007/978-3-030-61377-8_46

[40] Z. Shaheen, G. Wohlgenannt, E. Filtz, Large Scale Legal Text Classification Using Transformer Models, *arXiv preprint*, arXiv: 2010.12871, October, 2020. https://doi.org/10.48550/arXiv.2010.12871

[41] D. Liga, L. Robaldo, Fine-tuning GPT-3 for Legal Rule Classification, *Computer Law & Security Review*, Vol. 51, Article No. 105864, November, 2023. https://doi.org/10.1016/j.clsr.2023.105864

[42] L. Wan, G. Papageorgiou, M. Seddon, M. Bernardoni, Long-length Legal Document Classification, *arXiv preprint*, arXiv: 1912.06905, December, 2019. https://doi.org/10.48550/arXiv.1912.06905

[43] J. Lee, Legal Text Classification in Korean Sexual Offense Cases: From Traditional Machine Learning to Large Language Models with XAI Insights, *Artificial Intelligence and Law*, pp. 1-22, May, 2025. https://doi.org/10.1007/s10506-025-09454-w

[44] M.-Y. Kim, J. Rabelo, K. Okeke, R. Goebel, Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods, *The Review*

of Socionetwork Strategies, Vol. 16, No. 1, pp. 157-174, April, 2022.
https://doi.org/10.1007/s12626-022-00103-1

[45] Y. Aoki, M. Yoshioka, Y. Suzuki, Data-Augmentation Method for BERT-Based Legal Textual Entailment Systems in COLIEE Statute Law Task, *The Review of Socionetwork Strategies*, Vol. 16, No. 1, pp. 175-196, April, 2022.
https://doi.org/10.1007/s12626-022-00104-0

[46] M. Fujita, T. Onaga, A. Ueyama, Y. Kano, Legal Textual Entailment Using Ensemble of Rule-Based and BERT-Based Method with Data Augmentation by Related Article Generation, *JSAI International Symposium on Artificial Intelligence*, Kyoto, Japan, 2022, pp. 138-153.
https://doi.org/10.1007/978-3-031-29168-5_10

[47] H. Chu, H. Chu, P. Nguyen, M. Nguyen, Causal Relation-Aware Data Augmentation for Legal Textual Entailment, *30th International Conference on Applications of Natural Language to Information Systems*, Kanazawa, Japan, 2025, pp. 396-410.
https://doi.org/10.1007/978-3-031-97141-9_27

[48] S. Liu, J. Cao, Y. Li, R. Yang, Z. Wen, Low-resource Court Judgment Summarization for Common Law Systems, *Information Processing & Management*, Vol. 61, No. 5, Article No. 103796, September, 2024.
https://doi.org/10.1016/j.ipm.2024.103796

[49] A. Deroy, K. Ghosh, S. Ghosh, Applicability of Large Language Models and Generative Models for Legal Case Judgement Summarization, *Artificial Intelligence and Law*, Vol. 33, No. 4, pp. 1007-1050, December, 2025.
https://doi.org/10.1007/s10506-024-09411-z

[50] I. Benedetto, L. Cagliero, M. Ferro, F. Tarasconi, C. Bernini, G. Giacalone, Leveraging Large Language Models for Abstractive Summarization of Italian Legal News, *Artificial Intelligence and Law*, pp. 1-21, February, 2025.
https://doi.org/10.1007/s10506-025-09431-3

[51] W. Ansar, S. Goswami, A. Chakrabarti, Lesum: Cost-Effective LLM-Driven Hybrid Summarization of Indian Legal Judgments, Available at SSRN 5314737, June, 2025.
https://dx.doi.org/10.2139/ssrn.5314737

[52] H. Mentzingen, N. António, F. Bacao, Effectiveness in Retrieving Legal Precedents: Exploring Text Summarization and Cutting-Edge Language Models Toward a Cost-Efficient Approach, *Artificial Intelligence and Law*, pp. 1-21, February, 2025.
https://doi.org/10.1007/s10506-025-09440-2

[53] P. Italiani, G. Moro, L. Ragazzi, Enhancing Legal Question Answering with Data Generation and Knowledge Distillation from Large Language Models, *Artificial Intelligence and Law*, pp. 1-26, July, 2025.
https://doi.org/10.1007/s10506-025-09463-9

[54] F. Sovrano, M. Palmirani, S. Sapienza, V. Pistone, DiscoLQA: Zero-shot Discourse-based Legal Question Answering on European Legislation, *Artificial Intelligence and Law*, Vol. 33, No. 2, pp. 323-359, June, 2025.
https://doi.org/10.1007/s10506-023-09387-2

[55] Y. Wang, X. Shen, Z. Huang, L. Niu, S. Ou, cLegal-QA: A Chinese Legal Question Answering with Natural Language Generation Methods, *Complex & Intelligent Systems*, Vol. 11, No. 1, Article No. 77, January, 2025.
https://doi.org/10.1007/s40747-024-01675-x

[56] K. Veningston, A. Mishra, Dataset for Legal Question Answering System in the Indian Judiciary Context, *Data in Brief*, Vol. 60, Article No. 111647, June, 2025.
https://doi.org/10.1016/j.dib.2025.111647

[57] X. Wang, X. Zhang, V. Hoo, Z. Shao, X. Zhang, LegalReasoner: A Multi-stage Framework for Legal Judgment Prediction via Large Language Models and Knowledge Integration, *IEEE Access*, Vol. 12, pp. 166843-166854, November, 2024.
https://doi.org/10.1109/ACCESS.2024.3496666

[58] Y. Le, S. Xiao, Z. Xiao, K. Li, Topology-aware Multi-task Learning Framework for Civil Case Judgment Prediction, *Expert Systems with Applications*, Vol. 238, Part F, Article No. 122103, March, 2024.
https://doi.org/10.1016/j.eswa.2023.122103

[59] B. Wei, Y. Yu, L. Gan, F. Wu, An LLMs-based Neuro-symbolic Legal Judgment Prediction Framework for Civil Cases, *Artificial Intelligence and Law*, pp. 1-35, February, 2025.
https://doi.org/10.1007/s10506-025-09433-1

[60] D. Trautmann, A. Petrova, F. Schilder, Legal Prompt Engineering for Multilingual Legal Judgement Prediction, *arXiv preprint*, arXiv: 2212.02199, December, 2022.
https://doi.org/10.48550/arXiv.2212.02199

[61] M. Masala, R. C. A. Iacob, A. S. Uban, M. Cidota, H. Velicu, T. Rebedea, M. Popescu, JurBERT: A Romanian BERT Model for Legal Judgement Prediction, *Natural Legal Language Processing Workshop 2021*, Punta Cana, Dominican Republic, 2021, pp. 86-94.
https://doi.org/10.18653/v1/2021.nllp-1.8

[62] J. Sun, S. Huang, C. Wei, Chinese Legal Judgment Prediction via Knowledgeable Prompt Learning, *Expert Systems with Applications*, Vol. 238, Part E, Article No. 122177, March, 2024.
https://doi.org/10.1016/j.eswa.2023.122177

[63] W. Su, B. Yue, Q. Ai, Y. Hu, J. Li, C. Wang, K. Zhang, Y. Wu, Y. Liu, JUDGE: Benchmarking Judgment Document Generation for Chinese Legal System, *48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Padua, Italy, 2025, pp. 3573-3583.
https://doi.org/10.1145/3726302.3730295

[64] Q. Li, Q. Zhang, Court Opinion Generation from Case Fact Description with Legal Basis, *2021 AAAI Conference on Artificial Intelligence*, Virtual Conference, 2021, pp. 14840-14848.
https://doi.org/10.1609/aaai.v35i17.17742

[65] H. Li, J. Ye, Y. Hu, J. Chen, Q. Ai, Y. Wu, J. Chen, Y. Chen, C. Luo, Q. Zhou, Y. Liu, Casegen: A Benchmark for Multi-Stage Legal Case Documents Generation, *arXiv preprint*, arXiv: 2502.17943, February, 2025.
https://doi.org/10.48550/arXiv.2502.17943

[66] A. Ivaschenko, O. Golovnin, I. Syusin, A. Krivosheev, M. Aleksandrova, Ontology Based Text Understanding and Text Generation for Legal Technology Applications, *Science and Information Conference*, London, United Kingdom, 2023, pp. 1080-1089.
https://doi.org/10.1007/978-3-031-37963-5_75

[67] X. Li, L. Huang, Y. Zhou, C. Shao, TST-GAN: A Legal Document Generation Model Based on Text Style Transfer, *4th International Conference on Robotics, Control and Automation Engineering*, Wuhan, China, 2021, pp. 90-93.
https://doi.org/10.1109/RCAE53607.2021.9638919

[68] W. Huang, X. Liao, Z. Xie, J. Qian, B. Zhuang, S. Wang, J. Xiao, Generating Reasonable Legal Text Through the Combination of Language Modeling and Question Answering, *29th International Joint Conference on Artificial Intelligence*, Yokohama, Japan, 2021, pp. 3687-3693.

[69] R. Chikkamath, V. R. Parmar, Y. Otiefy, M. Endres. Patent classification using bert-for-patents on USPTO,

5th International Conference on Machine Learning and Natural Language Processing, Sanya, China, 2022, pp. 20-28.
https://doi.org/10.1145/3578741.3578746

[70] J. Niklaus, V. Matoshi, P. Rani, A. Galassi, M. Stürmer, I. Chalkidis, Lextreme: A multi-lingual and multi-task benchmark for the legal domain, *arxiv preprint*, arXiv: 2301.13126, January, 2023.
https://doi.org/10.48550/arXiv.2301.13126

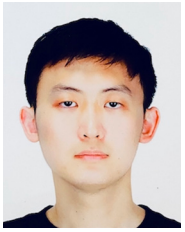
[71] I. Chalkidis, M. Fergadiotis, N. Manginas, E. Katakalou, P. Malakasiotis, Regulatory Compliance Through Doc2Doc Information Retrieval: A Case Study in EU/UK Legislation Where Text Similarity Has Limitations, *arXiv preprint*, arXiv: 2101.10726, January, 2021.
https://doi.org/10.48550/arXiv.2101.10726

[72] Y. Ma, Y. Shao, Y. Wu, Y. Liu, R. Zhang, M. Zhang, S. Ma, LeCaRD: a legal case retrieval dataset for Chinese law system, *44th international ACM SIGIR conference on research and development in information retrieval*, Virtual Event, Canada, 2021, pp. 2342-2348.
https://doi.org/10.1145/3404835.3463250

[73] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, J. Ge, Lawbench: Benchmarking legal knowledge of large language models, *arxiv preprint*, arXiv: 2309.16289, September, 2023.
https://doi.org/10.48550/arXiv.2309.16289

[74] Z. Chen, P. Ren, F. Sun, X. Wang, Y. Li, S. Zhao, T. Yang, SLARD: A Chinese Superior Legal Article Retrieval Dataset, *31st International Conference on Computational Linguistics*, Abu Dhabi, UAE, 2025, pp. 740-754.
https://aclanthology.org/2025.coling-main.50/

Biographies



Peter Jingzhou Lai, assistant professor at the School of Software Engineering, Quanzhou University of Information Engineering, China.

Ling Xia Liao, full professor in the School of Artificial Intelligence, Guilin University of Aerospace Technology, China.

Jie Chen, graduate student in the School of Electronic Engineering and Automation, Guilin University of Electronic Technology, China.

Miao Zhang, professor at Quanzhou University of Information Engineering, China.

Han-Chieh Chao, full professor in Fo Guang University, national Donghua University, Tamkang University, Taiwan, and UCSI University, Malaysia.