A Survey on Computational Law in the Era of Large Language Models

Anonymous ACL submission

Abstract

Research on computational law aims to build a bridge between Computer Science and Law. The application of AI techniques to the field of Law (AI for Law), as well as the regulation of issues arising from LLMs (Law for AI), have become important topics attracting significant attention from both AI researchers and legal professionals. AI technology, when applied effectively, has the potential to free legal professionals from the burden of repetitive tasks. At the same time, with appropriate policy support from these professionals, AI can be guided to evolve in a safer direction. This paper presents an overview of the development of computational law research from both "AI for Law" and "Law for AI" perspectives. We carry out experiments, interviews and provide a comprehensive analysis of current works, paving the way for future exploration. Detailed information regarding our work can be found at https://anonymous.4open.science/r/LLM-Regulation-E813/README.md.

1 Introduction

The emergence of large language models (LLMs) presents new opportunities for the development of computational methodologies in many social science fields. However, researchers often fail to understand how to conduct interdisciplinary research when they are unfamiliar with another field of knowledge. This is especially the case in the field of law, which is highly knowledge-intensive and has a high threshold for entry. The advent of LLMs brings new opportunities for many tasks in the legal field (Su et al., 2024; Li et al., 2024a; Gao et al., 2024; Chen et al., 2024), while also introducing numerous new legal issues (Wang et al., 2023a). Doing research on these issues requires a detailed understanding of both the computer science and legal fields. This survey aims to provide an overview of existing research and guidance for future studies for researchers who want to conduct computational law research in the era of LLMs.

Specifically, in this paper, we divide the research of computational law into two primary areas: AI for Law and Law for AI. It answers two questions: How can LLMs support legal tasks, and how should the risks brought by LLMs be regulated by law?

In the AI for Law section, this paper mainly addresses two issues:

1) Incorporating legal knowledge into LLMs. At present, many LLMs are applied to tasks in the legal field. In particular, the superior generation capabilities of LLMs make many previously impossible legal intelligence tasks possible. However, general-domain LLMs lack the necessary legal knowledge and therefore cannot accurately complete legal AI tasks on their own. Therefore, how to incorporate legal knowledge into the model has become an important issue. In this paper, we discuss how to build legal LLMs from the perspectives of base models and training methods.

2) Application of LLMs in legal professional works. In real life, the application of law varies among different groups. For example, lawyers often represent the interests of their clients, while judges should focus more on fairness. Students are more engaged in theoretical research, whereas the public has a greater need for legal assistance. This paper analyzes the application of LLMs to legal business from the perspectives of judges' work, lawyers' work, law school teaching, and legal aid.

In the Law for AI section, this paper mainly focuses on how to regulate the risks brought by LLMs through rules. As the capabilities of LLMs increase, their applications in daily life could bring a series of risks. Based on a comprehensive literature review, we categorizes the LLM risks into 18 types and summarizes existing AI regulations worldwide.

2 AI For Law

LLMs have demonstrated commendable proficiency across a multitude of downstream applications (Zhang et al., 2024a; Anand et al., 2023; Zhuang et al., 2023). In the realm of law, traditional AI models have struggled to distinguish between legal language and everyday language. For instance, the word "consideration" means "the price paid for a promise" in contract law, which is quite different from its ordinary meaning. However, with LLMs, machines can now comprehend legal language more accurately, enabling their application in a variety of legal scenarios such as contract drafting (Lam et al., 2023), judgment prediction (Feng et al., 2022), and similar case retrieval (Feng et al., 2024).

Despite the potential of LLMs in the legal domain, they have several limitations. Issues such as hallucinations (Dahl et al., 2024; Magesh et al., 2024) and delays in knowledge updates (Padiu et al., 2024), are still important problems that have not been solved. To mitigate these issues and truly integrate LLMs into legal practice, human involvement remains crucial. This then raises the questions: How to incorporate legal knowledge into LLMs? How can LLMs contribute to the legal field with the assistance of legal professionals?

2.1 Legal LLMs

In this section, we survey legal LLMs in recent years, and as shown in the Appendix A, a large number of legal LLMs have emerged in the past few years. We briefly introduce existing legal LLMs, their training processes, and the training datasets. As illustrated in Figure 1, existing legal LLMs are trained based on the general LLMs, and be subsequently pre-trained, fine-tuned, and retrieval-enhanced. We elaborate on the details of each module in this section.

2.1.1 Base Model

As shown in Figure 1, most existing legal LLMs are fine-tuned based on general LLMs, such as LLaMA (Touvron et al., 2023), GPT (Brown et al., 2020), Bloom (Le Scao et al., 2022), and GLM (Du et al., 2021). There are some legal LLMs with open-source code based on a variety of base models, such as the LawGPT (Zhou et al., 2024) series models, which support general Chinese base models like Chinese-LLaMA (Cui et al., 2023b) and ChatGLM (GLM et al., 2024). Although new base models are released frequently, it is simple to switch to the latest base models to ensure the knowledge remains up-to-date.

2.1.2 Legal Corpus

Legal datasets play a crucial role both from the perspective of training LLMs and evaluating them. As shown in the Appendix C, there are a vast number of existing Legal AI datasets. These datasets contain legal texts drawn from official and unofficial sources. Official data sources are the bedrock of legal LLMs, comprising statutory laws, judicial decisions, legal textbooks, and scholarly articles. Unofficial data sources include resources such as legal question-and-answer datasets from online forums. They are particularly valuable in understanding how legal principles and requirements are applied in real-world scenarios.

The efficacy of legal LLMs is deeply contingent on the quality and authority of their training data. The models must adeptly address ethical and legal issues, including bias mitigation and adherence to legal standards, to ensure their responsible use and effectiveness in legal practices.

2.1.3 Training Process

Most legal LLMs are built on base models and are trained using legal datasets following the "pre-training, fine-tuning, and retrieval enhancement" processes. For legal LLMs, the continual pre-training process aids in understanding legal language, the fine-tuning process helps to achieve a more precise comprehension of legal language and grasp the legal logic within it, and the RAG (Retrieval-Augmented Generation) contributes to the model's ability to provide answers based on the precise legal knowledge it has retrieved. Details can be found in Appendix G.

Continual Pre-training. Continual Pre-training refers to the process of further training the model on a large-scale unlabelled dataset (Ji et al., 2023a; Cui et al., 2023c) based on the pre-trained base model, with the aim of enhancing the model's performance in the comprehension of legal texts.

Fine-tuning. Fine-tuning refers to the process of making precise adjustments to the pre-trained model using a small-scale labeled dataset, enabling the model to adapt better to specific tasks (Min et al., 2023).

In the fine-tuning of legal LLMs, researchers typically direct the existing general-domain LLMs to perform supervised fine-tuning (SFT) on le-

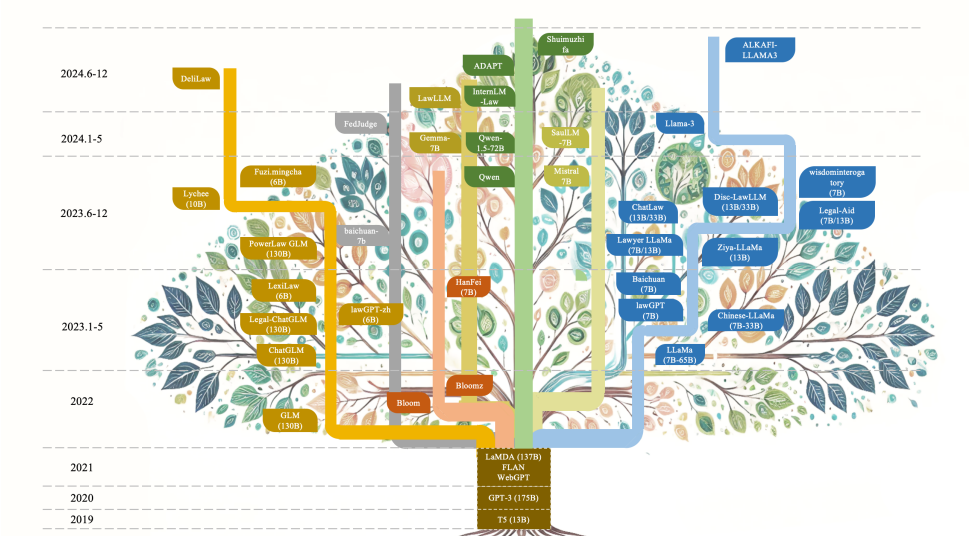


Figure 1: Legal LLM Tree.

gal datasets, such as legal documents, legal articles, and high-quality legal question-and-answer datasets (Cui et al., 2023a; Haitao Li, 2024; Huang et al., 2023b).

Retrieval-Augmentation. General domain LLMs often suffer from hallucination (Ji et al., 2023b; Huang et al., 2023a). In the legal field, the content generated by LLMs needs to be highly knowledgeable and reliable. However, the hallucination issue can lead to the creation of fabricated legal provisions or falsifyied legal facts, rendering LLMs unreliable (Magesh et al., 2024). At the same time, legal knowledge is continuously updated. If we simply rely on pre-training and fine-tuning LLMs, the models may remember and keep forever the outdated knowledge learned in the corpus used for training. To address these issues, many researchers have optimized legal LLMs using the retrieval-augmented approach (Cui et al., 2023a; Huang et al., 2023b; Cui et al., 2024). Retrieval-augmentation is akin to giving the LLM an open-book exam, where the most relevant corpus from the retrieval library is inputted into the LLM, allowing it to reference the retrieved content for its response.

2.2 Evaluation Benchmark

2.2.1 Tasks and Metrics

Evaluation of legal LLMs is generally structured around the feature of the downstream tasks. Tasks can be divided into two categories (Details can be found in Appendix H):

(1) Generation Tasks: These tasks require models to generate text (e.g., summarization or legal reasoning). The quality of generated outputs is typically assessed using metrics such as ROUGE-L (Steffes et al., 2023; Mullick et al., 2022) and BERT-Score (Kumar et al., 2024; Benedetto et al., 2023; Joshi et al., 2024). With the advancement of LLMs' comprehension capabilities, some evaluation tasks have begun incorporating LLMs as judges to assess performance from multiple dimensions (Cui et al., 2023a; Li et al., 2024c).

(2) Decision Tasks: These tasks involve classification or extraction, such as multiple-choice legal question answering (Pahilajani et al., 2024), legal case retrieval (Feng et al., 2024; Padiu et al., 2024), or judgment prediction (Wu et al., 2023; Wang et al., 2024). In this setting, evaluation metrics often include Recall, Accuracy, F1 scores, Mean Absolute Error (MAE). The detailed descriptions and calculation methods for the metrics are shown in Appendix F.

In addition, some evaluation frameworks[1] incorporate supplementary tasks assessing model safety and performance (System et al., 2023)

2.2.2 Benchmarks

Existing legal LLM benchmarks are essential for evaluating model performance across legal tasks. Comprehensive benchmarks encompass diverse generative or decision tasks, often classified by task type (Fei et al., 2023) or cognitive ability, such as memory, reasoning, and ethics (Fei et al., 2023). Given the importance of logic in law, recent bench-

[1] https://data.court.gov.cn/pages/modelEvaluation.html

marks increasingly focus on assessing LLMs' legal reasoning capabilities (Guha et al., 2024; Dai et al., 2023). Details can be found in Appendix E.

LawBench (Fei et al., 2023) assesses LLMs across three cognitive levels: memorization, understanding, and application. It includes 20 diverse tasks, making it more aligned with real-world legal applications than multiple-choice-based evaluations. Results are shown in Figure 4.

LegalBench (Guha et al., 2024) takes an interdisciplinary approach, consisting of 162 tasks covering six types of legal reasoning. Developed with contributions from legal professionals, it aligns with legal reasoning frameworks and provides a structured evaluation of LLMs' ability to perform legally relevant reasoning tasks. Results are shown in Table 5.

LexEval (Li et al., 2024b) is the largest Chinese legal benchmark to date, featuring 23 tasks and 14,150 questions. It introduces a new taxonomy of legal cognitive abilities and evaluates not only fundamental legal knowledge but also ethical concerns in legal AI applications. Results are shown in Table 6 and Table 7.

LAiW (Dai et al., 2023) is the first Chinese legal LLM benchmark grounded in the logic of legal practice. It structures legal reasoning into three levels-basic information retrieval, legal foundation inference, and complex legal application-aligning with syllogistic legal reasoning. The overall scores of legal capability of LLMs are ranked as Figure 5.

2.3 Applications

Legal LLMs are designed to assist human in accomplishing legal tasks. As shown in Section 2.2, different LLMs perform differently on legal tasks. Therefore, there is no single best legal LLM; each model LLM has its strengths and weaknesses when applied to legal tasks. Consequently, legal professionals need to play a crucial role in leveraging these models in legal scenarios. In this section, we identify the beneficiaries of legal LLMs and discuss the potential usages and issues of LLMs in practice. Details can be found in Appendix I.

2.3.1 Applications to Judges

LLMs can support judges' work, such as improving trial efficiency, alleviating the pressure on judges' work, and promoting fairness and justice. However, when LLMs come into use in courts, it will face many practical challenges. These include issues such as how to deal with public trust in LLMs, whether judges can apply LLMs to promote good governance, and how to avoid biases and inequalities. A study indicates that decision-makers using AI, while not inclined to automatically follow algorithmic suggestions, may exhibit "selective adherence" when these suggestions align with their stereotypes (Alon-Barkat and Busuioc, 2023). This tendency to selectively adopt suggestions could negatively impact citizens who are already in a disadvantaged position.

To the dilemmas of applying LLMs in judicial activities, it is necessary to develop innovative, reliable, and secure legal LLMs. In fact, the topic of using AI in public governance has been widely discussed, and some governments and organizations have now issued some policies and recommendations, which may in the future become the basis for solving the problems of judges applying LLMs (onAI Good Governance , OxCAIGG). For example, the Supreme People's Court of the People's Republic of China[2] and the UK Courts and Tribunals Judiciary[3] both emphasize that AI still serves as an auxiliary tool in judges' work. It should be stressed that these documents only hold guiding significance, and further implementation at the legal level is still pending in the future.

2.3.2 Applications to Lawyers

LLMs have reshaped the way of understanding and acquiring knowledge, bringing significant benefits to the work of lawyers in various professional levels and business area. In addition to improving the efficiency of lawyers, factors such as client demands and trust may lead to the inevitable choice of using LLMs for lawyers.

LLMs can be used for many tasks such as contract review, due diligence, document draft, case summary, cross-examine questions, evidence strengthen suggestions, etc. (Perlman, 2023; Tan et al., 2023; Noonan, 2023), enabling lawyers to concentrate on core tasks and deliver more cost-effective solutions to their clients. Currently, a significant number of law firms and legal tech companies have publicly announced their integration of the LLM for specific use cases (Shaver, 2024; Iu and Wong, 2023).

According to the characteristics of lawyers' work, we categorize the current applications for

[2] https://www.chinacourt.org/article/detail/2022/12/id/7057666.shtml

[3] https://www.judiciary.uk/wp-content/uploads/2023/12/AI-Judicial-Guidance.pdf

lawyers into three parts: common functions, functions for litigation lawyers, and functions for non-litigation lawyers (Figure 7). It should be noted that the categorization here is primarily for the convenience of discussion, as the identity of a lawyer may not be unique in practice.

First, basic applications refer to the functions commonly used by all lawyers in their professional activities. LLMs have demonstrated good performance in tasks such as drafting, translation, review, summarization, polishing, etc. However, there may be certain issues in LLMs outputs. For instance, when analyzing bona fide acquisition system, ChatGPT believes someone is unable to obtain the ownership if he/she fails to return the subject matter maliciously which have been bought in good faith. It is evidently inconsistent with legal provisions. Thus, the outputs of LLMs require careful evaluation by lawyers to ensure accuracy and reliability.

Second, applications for litigation lawyers include AI moot court, referee result prediction, evidence strengthen suggestions, etc. Litigation lawyers typically engage in activities such as litigation and arbitration, involving tasks such as client meetings, evidence collection, and court debates. The mentioned applications can provide convenience for their work. Take AI moot court as an example, AI can effectively assist the lawyer in quickly identifying potential shortcomings in the regulations, evidence, argument points, and debate strategies they have prepared, thereby increasing the likelihood of a successful outcome.

Third, LLMs can provide support for non-litigation lawyers in due diligence, compliance risk assessment, legal advisor assistant, and other tasks. For instance, IP lawyers conducting Freedom to Operate (FTO) analysis need to comprehensively evaluate existing patents in a specific technical field to help businesses avoid unintentional infringement of others' patent rights. Lawyers often spend a significant amount of time on FTO analysis as technical terms are hard to comprehend and there's a lot of document data to compare. Obviously, AI can play a huge helpful role in these two aspects. In the future, LLMs may further enhance its assistance in FTO analysis by enabling batch import of patent files, using patent files as training data for post-training or fine-tuning, subdividing patent technologies, and automating the generation of analysis reports.

Overall, there are many LLM applications for lawyers, many of which have already performed well. For lawyers, as LLMs become more sophisticated and capable of performing complex legal tasks, they need to consider assigning simple, repetitive tasks to LLMs and focus more on areas where they can add the most value. Besides, lawyers also need to be vigilant about potential issues such hallucinations, copyright infringement, privacy breaches, discrimination, etc., when using LLMs.

2.3.3 Applications to Law School Students

LLM as a research subject. LLMs significantly reduce AI accessibility barriers for law students.

Prior to LLMs, while many law schools discuss introducing courses on AI and its legal implications (Goldsworthy, 2020), the study LAW2020[4] shows that only 20% law students have adopted such programs. Earlier legal AI systems like LEGAL-BERT (Chalkidis et al., 2020) and Lawformer (Xiao et al., 2021) required technical expertise for tasks, such as vector encoding and similarity calculation, training and deployment of models, and so on.

LLMs' intuitive interface enables non-technical students to engage with advanced technology through prompt engineering. With some basic knowledge of LLMs, law students can engage with LLMs and analyze the responses from a legal perspective. After mastering the fundamentals of LLMs, law students can provide professional support for policy-making, AI ethics evaluation, and intelligent legal governance.

LLM as a learning tool. In the past, specialized models had to be individually designed for each task. However, a legal LLM can accommodate a broad spectrum of learning requirements, such as generating summaries and extracting legal elements (Kasneci et al., 2023). When providing case briefs, LLMs not only extract legal facts and conclusions, but also help analyze judgment logic and the sentencing factors (de Faria et al., 2024; Yue et al., 2023). Currently, attempts have been made to introduce LLMs into the law school classroom. Lexis+AI (Mika, 2022), a tool that supports conversational search, intelligent legal drafting, insightful summarization, and document analysis, is set to be accessible to 100,000 law students in the 2024 spring semester[5].

[4] https://abovethelaw.com/law2020/cognifying-legal-education/

[5] https://www.abajournal.com/web/article/

While LLMs enhance legal education through interactive feedback and analytical support, their outputs require professional verification. Commercial implementations like Lexis+AI show practical adoption, though currently restricted to advanced students due to residual hallucination risks. In addition, an over-reliance on LLMs could potentially hinder students' growth and suppress their creativity, necessitating guided usage with clear risk disclosures regarding accuracy and bias.

2.3.4 Applications to the General Public

Accessing legal information is challenging for the general public. Financial constraints, fear of the law, and complex procedures often hinder the average person from seeking legal advice. Despite government's efforts to expand public legal aid, these measures often fail to meet individual needs (Mansfield and Trubek, 2011).

Before the advent of LLMs, digital tools were used to bridge the gap between ordinary people and lawyers. These tools are based on rules or incorporate traditional AI techniques. To some extent, they reduce the cost of money when people seek legal help and expand the coverage of legal aid. Additionally, they allow users to ask questions without psychological burden, especially for private matters concerning marriage and family. However, these tools have limitations regarding their capabilities and coverage.

Rule-based legal tools. These tools, designed by legal experts, aim to provide precise calculations for specific legal requirements, but face challenges in reusability and accessibility for the average user.

Traditional AI-based legal tools. These tools are typically AI combined with rules (Dias et al., 2022)[6]. Due to the limitation of model scale, they can only serve for certain services, and perform poorly on these limited applications. Those tools suffer from lack of service diversity and accuracy.

LLM-based legal tools. LLMs combine the accessibility of traditional legal tools with improved accuracy and adaptability. Their ability to translate legal jargon into plain language lowers comprehension barriers for non-specialists. Extensive pre-training data enables LLMs to offer reasonably accurate responses based on their experience, even in unfamiliar situations.

With a rich pre-trained corpus and external knowledge base or tools, LLMs can supply abundant supporting material[7]. The average person, for example, may only have a vague awareness of potential legal risks, without understanding how to address them. Legal aid LLMs, such as Shuimuzhifa[8], can restructure the case based on the user's description, clearly inform the user of the current legal situation, provide subsequent coping strategies (such as rights protection or prosecution), and match a template to assist the user in writing legal documents.

3 Law for AI

While AI empowers the legal field, it also generates security risks in dimensions such as training data, algorithms, and generated content. In the aspect of "Law for AI", many positive strategies have emerged at the international level. However, many lawsuits under trailed indicate that the legal issues triggered by LLM risks still have a significant impact on the legal system. How to strike a balance among rights protection, technological development, and national security has become an inevitable key issue.

This part will conduct an in-depth discussion on LLM risks governance and core legal issues. Section 3.1 reviews global AI governance documents, LLM risks, and governance perspectives. Section 3.2 discusses three key legal issues related to LLMs, along with the current state of research and practice.

3.1 LLM Risk Governance

Based on a review of 50+ global legal documents concerning LLMs (including regulations, standards, guidelines, policy documents, research reports, forum initiatives, and other related materials, referring to Appendix B), we find that, in response to the challenges posed by LLMs in areas such as security, copyright, privacy, and ethics, the framework for governing LLM risks has gradually become clearer, and the basis for governance has transitioned from abstract policies and initiatives to legally binding laws and regulations.

3.1.1 LLM Risks and Risks Mapping

As Figure 2 shows, we categorize 18 types of LLM risks into four aspects. Details can be found in Appendix D.

Then we map these risks to the relevant documents (Figure 3) and find that, on a global scale,

law-students-gain-access-to-lexis-ai-generative-ai

[6] http://www.12348.gov.cn/#/homepage

[7] https://tongyi.aliyun.com/farui

[8] https://www.shuimuzhifa.com/

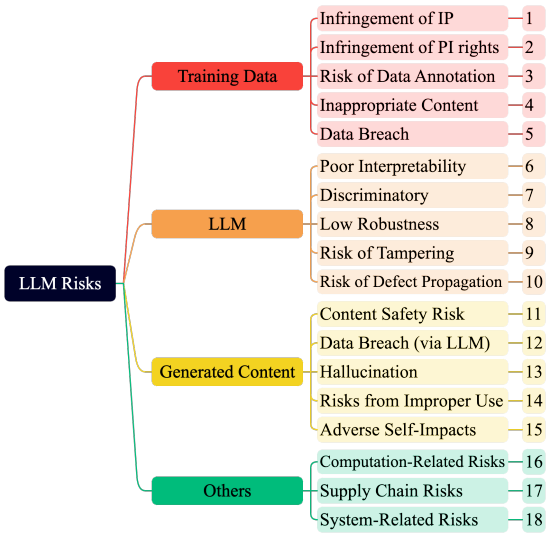


Figure 2: LLM Risks

the top three LLM risks receiving the most attention are Generated Content – Risks from Improper Use; Training Data – Inappropriate Content; and Generated Content – Content Safety Risk. The three risks that receive the least attention are Supply Chain Risks, Risk of Defect Propagation, and Computation-Related Risks.

3.1.2 LLM Risks Governance Approaches

Additionally, from the perspective of risk governance, current regulations primarily encompass three governance approaches.

The first is a lifecycle-based governance perspective for LLMs, which involves managing risks throughout the entire lifecycle of an LLM, including design, development, training, testing, deployment, use, and maintenance.

The second approach is a risk source-based governance perspective, which involves identifying and managing risks based on their specific sources.

The third typical governance perspective involves establishing security requirements specific to LLM service scenarios.

3.2 LLM related core legal issues

Although the above risks originate from different sources, they are not completely independent. Two typical legal issues, copyright infringement of training data and disputes over personal information rights, are both caused by risks during the training phase. However, they only surface during the content generation phase after the generation of incorrect content, thus leading to risks associated with the generated content.

Additionally, the copyrightability of generated content has received widespread attention in the field of copyright. Although this issue does not fall under security risks, it presents new challenges to the copyright system. Therefore, this paper will also discuss this issue.

3.2.1 Copyright Infringement of Training Data

The training process of an LLM relies on a large amount of data. If copyrighted works are used without authorization, it may lead to the risk of copyright infringement, and such cases are not uncommon.

To reduce the development costs for LLM developers, some scholars have proposed the doctrines of "nonexpressive use" (Sag, 2023; Flynn et al., 2020) or "temporary reproduction" (Wang Qian, 2024), arguing that the use of works during the training process does not constitute copyright infringement. Other scholars have put forward the fair use principle, which allows the unauthorized use of copyrighted works under specific conditions.

At the legislative level, different countries have different provisions regarding fair use. For example, the United States adopts the "four-factor test"[9] to provide space for interpreting the fair use of training data. The EU employs mechanisms of implied consent from authors to regulate the fair use of data in "text and data mining". However, Article 24 of China's Copyright Law provides an enumerative list for fair use, making it difficult to include the use of training data within the scope.

We contends that, in addition to the fair use doctrine, approaches such as constructing training data licensing platforms, obtaining unified authorization through collective copyright management organizations, and establishing automated licensing mechanisms should be considered from the perspective of facilitating authorizations for copyright holders. Whether through institutional breakthroughs or operational optimizations, the issue of copyright infringement in training data must be adequately addressed.

3.2.2 Copyrightability of Generated Content

The issue of copyrightability for AI-generated content is another contentious hotspot for LLMs in the field of intellectual property. Firstly, on the premise

[9]The US Copyright Law, Article 107.

that "AI itself cannot be regarded as an author in the context of copyright law", the ownership of copyright remains controversial due to different perspectives. Secondly, the legal provision that the author in copyright law does not include "non-human entities" is often used in U.S.[10][11] cases to deprive AI-generated content of the eligibility for copyright protection. In contrast, Chinese cases[12][13] regard whether the generated content meets the originality requirement as the core criterion for copyright protection. The originality of a work is determined by judging the degree of the parties' intellectual contribution in the content generation process.

We argue the fact that AI used in the generation process does not inherently exclude the possibility of the generated content being protected by copyright. However, considering AI's highly automated capabilities, it is necessary to more rigorously assess on a case-by-case basis whether the generated content meets the originality requirements.

3.2.3 Personal Data Infringement

Personal data infringement typically arise when personal data used in training is obtained without the consent of the data subjects, leading to illegal processing or the leakage of personal data.

At the legal level, avoiding personal information infringement requires meeting the legality basis for data processing. The EU's General Data Protection Regulation (GDPR) and China's Personal Information Protection Law (PIPL) offer various methods, and the simplest approach among them is to obtain the consent of the data subject. To address the issue that data subjects have difficulty understanding privacy policies in practice, in addition to individual consent, regulatory authorities in many countries and regions have put forward stricter requirements for the use of personal data by LLMs. The response of the EDPB (EDPB, 2024) regarding "legitimate interests" in the Meta case serves as an example.

At the technical level, personal data anonymization can also meet the requirements for legal processing, but it is challenging to implement. In the relevant regulations of the GDPR and PIPL, anonymized personal data is no longer regarded as falling within the category of personal data. However, the legal standard for anonymization, which is "cannot identify a specific natural person in any way and cannot be re-identified"[14] is difficult to achieve. This is because methods such as database matching attacks still pose the possibility of identifying individuals (Narayanan and Shmatikov, 2008). Currently, the academic community has not reached a consensus on the feasibility of anonymization.

In the future, AI governance needs to develop further in a coordinated manner across multiple dimensions such as technology, law, and ethics. For example, enhancing data anonymization and model transparency through technological means; clarifying the legal basis for training data, the rights attributes of generated content, and the distribution of responsibilities through legislation; and guiding the sustainable development of AI technology through ethical norms. These measures will collectively promote the effective governance of LLM risks and lay the foundation for the safe development of AI.

4 Conclusion

In this paper, we introduce the history, current status, and future directions of computational law research. The capabilities of LLMs in the legal field have advanced to the point where they can assist legal professionals in completing some simple, repetitive tasks. However, there is still a significant room for improvement in areas such as how to infuse LLMs with legal knowledge, how to evaluate the capabilities of legal LLMs, and how to apply these models in legal practice. At the same time, the emergence of LLMs has brought about numerous legal issues, such as copyrightability and personal data infringement. This paper enumerates laws, regulations, and policy documents established by various countries and regions.

In the future, the development of policies in related fields will require contributions from more interdisciplinary researchers. We hope that this paper can serve as a guide for both AI researchers and legal professionals, providing a comprehensive picture of the current state of development. Based on this, researchers from various fields can gain a deeper understanding of other areas and conduct more interdisciplinary research.

[10] The US Copyright Office, Re: Zarya of the Dawn (Registration #VAu001480196)

[11] Thaler v. Perlmutter, No.22-1564 (D.D.C. August 18, 2023)

[12] Nanshan District People's Court, Shenzhen, Guangdong, China: Shenzhen Tencent v. Shanghai Yingxun, Case No. Y0305MC No.14010(2019).

[13] Beijing Internet Court, China: Li v. Liu, Case No. Y0491MC No.11279(2023).

[14] PIPL, Article 73.

Limitations

While we have struggled to conduct a comprehensive survey of AI & Law, several limitations remain. In the aspect of AI for Law, specific testing experiments have yet to be conducted on certain issues, such as the hallucination problem in large models and the application of legal knowledge graphs. In the aspect of Law for AI, it is challenging to cover all global regulations, policies, and cases, and as time progresses, new legal issues related to AI continue to emerge. Moving forward, we will continue to delve deeper into the technical challenges of AI for law and will periodically update related content on GitHub concerning law for AI.

Ethics Statement

The results presented in this survey are based on previously published articles or publicly accessible materials. All other sections are derived from the experiments conducted by the authors or their own viewpoints.

References

Benjamin Alarie, Anthony Niblett, and Albert H Yoon. 2018. How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, 68(supplement 1):106–124.

Vincent Aleven and Kevin D Ashley. 1997. Teaching case-based argumentation through a model and examples: Empirical evaluation of an intelligent learning environment. In *Artificial intelligence in education*, volume 39, pages 87–94. Citeseer.

Saar Alon-Barkat and Madalina Busuioc. 2023. Human–ai interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169.

Avishek Anand, Abhijit Anand, Vinay Setty, et al. 2023. Query understanding in the age of large language models. *arXiv preprint arXiv:2306.16004*.

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. Gentle: A genre-diverse multilayer challenge set for english nlp and linguistic evaluation. *arXiv preprint arXiv:2306.01966*.

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. Eur-lex-sum: A multi-and cross-lingual dataset for long-form summarization in the legal domain. *arXiv preprint arXiv:2210.13448*.

Irene Benedetto, Luca Cagliero, Francesco Tarasconi, Giuseppe Giacalone, and Claudia Bernini. 2023. Benchmarking abstractive models for italian legal news summarization. In *Legal Knowledge and Information Systems*, pages 311–316. IOS Press.

Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2022. Legal case document similarity: You need both network and text. *Information Processing & Management*, 59(6):103069.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tobias Brugger, Matthias Stürmer, and Joel Niklaus. 2023. Multilegalsbd: a multilingual legal sentence boundary detection dataset. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 42–51.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 19–28.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019b. Extreme multi-label legal text classification: A case study in eu legislation. *arXiv preprint arXiv:1905.10892*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. 2022. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *arXiv preprint arXiv:2203.07228*.

Open AI's Assistant ChatGPT and Andrew M. Perlman. 2022. The implications of openai's assistant for legal services and society. *SSRN Electronic Journal*.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. 2024. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023c. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. Laiw: A chinese legal large language models benchmark (a technical report). *arXiv preprint arXiv:2310.05620*.

Joana Ribeiro de Faria, Huiyuan Xie, and Felix Steffek. 2024. Automatic information extraction from employment tribunal judgements using large language models. *arXiv preprint arXiv:2403.12936*.

Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024. Enabling discriminative reasoning in LLMs for legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 784–796, Miami, Florida, USA. Association for Computational Linguistics.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.

João Dias, Pedro A Santos, Nuno Cordeiro, Ana Antunes, Bruno Martins, Jorge Baptista, and Carlos Gonçalves. 2022. State of the art in artificial intelligence applied to the legal domain. *arXiv preprint arXiv:2204.07047*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

EDPB. 2024. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of ai models. https://www.edpb.europa.eu/system/files/2024-12/edpb$_o pinion_2 02428_a i - models_e n.pdf$.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Maosong Cao, Fengzhe Zhou, Yining Li, Wenwei Zhang, Dahua Lin, et al. 2024. Internlm-law: An open source chinese legal large language model. *arXiv preprint arXiv:2406.14887*.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction: A survey of the state of the art. In *IJCAI*, pages 5461–5469.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485.

Sean Flynn, Christophe Geiger, João Pedro Quintais, Thomas Margoni, Matthew Sag, Lucie Guibault, and Michael W Carroll. 2020. Implementing user rights for research in the field of artificial intelligence: A call for international action. *Joint PIJIP/TLS Research Paper Series*, (48).

Cheng Gao, Chaojun Xiao, Zhenghao Liu, Huimin Chen, Zhiyuan Liu, and Maosong Sun. 2024. Enhancing legal case retrieval via scaling high-quality synthetic query-candidate pairs.

Ingo Glaser, Elena Scepankova, and Florian Matthes. 2018. Classifying semantic types of legal sentences: Portability of machine learning models. In *Legal Knowledge and Information Systems*, pages 61–70. IOS Press.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daniel Goldsworthy. 2020. The future of legal education in the 21st century. *TheADELAIDE LAW REVIEW*, 41(1):243–265.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in cjeu decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157.

Neel Guha, Daniel E Ho, Julian Nyarko, and Christopher Ré. 2022. Legalbench: Prototyping a collaborative benchmark for legal reasoning. *arXiv preprint arXiv:2209.06120*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish legalese language model and corpora. *arXiv preprint arXiv:2110.12201*.

Qian Dong Yiqun Liu Haitao Li, Qingyao Ai. 2024. Lexilaw: A scalable legal language model for comprehensive legal understanding.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023. Hanfei-1.0. https://github.com/siat-nlp/HanFei.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*.

Rebecca Howlett and Cynthia Sharp. 2023. Chatgpt: What lawyers need to know before using ai. *GP Solo eReport*, 12(11).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023b. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Kwan Yuen Iu and Vanessa Man-Yi Wong. 2023. Chatgpt by openai: The end of litigation lawyers? *Available at SSRN 4339839*.

Yunjie Ji, Yan Gong, Yong Deng, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023a. Towards better instruction following language models for chinese: Investigating the impact of training data and evaluation. *arXiv preprint arXiv:2304.07854*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. Ilur: Benchmark for indian legal text understanding and reasoning. *arXiv preprint arXiv:2407.05399*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Moniba Keymanesh, Micha Elsner, and Srinivasan Sarthasarathy. 2020. Toward domain-guided controllable summarization of privacy policies. In *NLLP@ KDD*, pages 18–24.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2022. Coliee 2022 summary: methods for legal document retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 51–67. Springer.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Hemanth Kumar, P Jayanth, et al. 2024. Large language models for indian legal text summarisation. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–5. IEEE.

Kwok-Yan Lam, Victor CW Cheng, and Zee Kin Yeong. 2023. Applying large language models for enhancing contract drafting. In *LegalAIIA@ ICAIL*, pages 70–80.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv e-prints*, pages arXiv–2211.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2024a. Legalagentbench: Evaluating llm agents in legal domain.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024b. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024c. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

Laura Manor and Junyi Jessy Li. 2019. Plain english summarization of contracts. *arXiv preprint arXiv:1906.00424*.

Marsha M Mansfield and Louise G Trubek. 2011. New roles to solve old problems: Lawyering for ordinary people in today's context. *NYL Sch. L. Rev.*, 56:367.

Antonio Mauricio, Vladia Pinheiro, Vasco Furtado, João Araújo Monteiro Neto, Francisco das Chagas Jucá Bomfim, André Câmara Ferreira da Costa, Raquel Silveira, and Nilsiton Aragão. 2023. Cdjur-br–a golden collection of legal document from brazilian justice with fine-grained named entities. *arXiv preprint arXiv:2305.18315*.

Karin Mika. 2022. Friend or foe? lexis artificial intelligence (ai) in legal writing. *Proceedings: Online Journal of Legal Writing Presentations*, 3(1):24.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, R Raghav, and Roshni Kar. 2022. An evaluation framework for legal document summarization. *arXiv preprint arXiv:2205.08478*.

Michael D Murray. 2023. Artificial intelligence and the practice of law part 1: Lawyers must be professional and responsible supervisors of ai. *Available at SSRN 4478588*.

Arvind Narayanan and Vitaly Shmatikov. 2008. Robust deanonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023a. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023b. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

Nick Noonan. 2023. Creative mutation: A prescriptive approach to the use of chatgpt and large language models in lawyering. *SSRN Electronic Journal*.

Oxford Commission onAl Good Governance (OxCAIGG). 2021. Ai in the public service: From principles to practice. https://oxcaigg.oii.ox.ac.uk/wp-content/uploads/sites/11/2021/12/AI-in-the-Public-Service-Final.pdf.

Bogdan Padiu, Radu Iacob, Traian Rebedea, and Mihai Dascalu. 2024. To what extent have llms reshaped the legal domain so far? a scoping literature review. *Information*, 15(11):662.

Anish Pahilajani, Samyak Rajesh Jain, and Devasha Trivedi. 2024. Nlp at uc santa cruz at semeval-2024 task 5: Legal answer validation using few-shot multiple-choice qa. *arXiv preprint arXiv:2404.03150*.

Andrew Perlman. 2023. The implications of chatgpt for legal services and society. *Mich. Tech. L. Rev.*, 30:1.

James Purtill. 2023. How chatgpt and other new ai tools are being used by lawyers, architects and coders. https://www.abc.net.au/news/science/2023-01-25/chatgpt-midjourney-generative-ai-and-future-of-work/101882580.

Rabee Qasem, Mohannad Hendi, and Banan Tantour. 2024. Alkafi-llama3: Fine-tuning llms for precise legal understanding in palestine. *arXiv preprint arXiv:2412.14771*.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. 2023. Scale: Scaling up the complexity for advanced language model evaluation. *arXiv preprint arXiv:2306.09237*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*.

Julien Rossi, Svitlana Vakulenko, and Evangelos Kanoulas. 2021. Verbcl: A dataset of verbatim quotes for highlight extraction in case law. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4554–4563.

Matthew Sag. 2023. Copyright safety for generative ai. *Hous. L. Rev.*, 61:295.

Robert Shaffer and Stephen Mayhew. 2019. Legal linking: citation resolution and suggestion in constitutional law. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 39–44.

Nicola Shaver. 2024. The use of large language models in legaltech. https://www.legaltechnologyhub.com/contents/the-use-of-large-language-models-in-legaltech/.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889.

Francesco Sovrano, Monica Palmirani, and Fabio Vitali. 2020. Legal knowledge extraction for knowledge graph based question-answering. In *Legal knowledge and information systems*, pages 143–153. IOS Press.

Bianca Steffes, Piotr Rataj, Luise Burger, and Lukas Roth. 2023. On evaluating legal summaries with rouge. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 457–461.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Quezi Bing, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. STARD: A Chinese statute retrieval dataset derived from real-life queries by non-professionals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10658–10671, Miami, Florida, USA. Association for Computational Linguistics.

Intelligent Judicial Technology Chief Engineer System, Zhejiang University, Shanghai Jiao Tong University, Ltd. Alibaba Cloud Computing Co., and iFLYTEK Research Institute. 2023. Evaluation metrics and assessment methods for large legal models (draft for comments). Technical report, Intelligent Judicial Technology Chief Engineer System. Accessed: 2025-02-06.

Jinzhe Tan, Hannes Westermann, and Karim Benyekhlef. 2023. Chatgpt as an artificial lawyer? In *AI4AJ@ ICAIL*.

Luke Taylor. 2023. Colombian judge says he used chatgpt in ruling. https://www.theguardian.com/technology/2023/feb/03/colombian-judge-chatgpt-ruling.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

Stefanie Urchs, Jelena Mitrovic, and Michael Granitzer. 2021. Design and implementation of german legal decision corpora. In *ICAART (2)*, pages 515–521.

Franjo Vučić. 2023. Changes in legal education in the digital society of artificial intelligence. In *International Conference on Digital Transformation in Education and Artificial Intelligence Application*, pages 159–176. Springer.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023b. Maud: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*.

Xuran Wang, Xinguang Zhang, Vanessa Hoo, Zhouhang Shao, and Xuguang Zhang. 2024. Legalreasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration. *IEEE Access*.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Zhu Chu Wang Qian. 2024. A preliminary exploration of the boundary between ai and copyright: Legal challenges and reflections under technological progress. *Chinese Editors Journal*, (08):56–62.

World Intellectual Property Organization (WIPO). 2023. World intellectual property indicators 2023. https://www.wipo.int/edocs/pubdocs/en/wipo-pub-941-2023-en-world-intellectual-property-indicators-2023.pdf.

Yang Wu, Masayuki Mukunoki, Takuya Funatomi, Michihiko Minoh, and Shihong Lao. 2011. Optimizing mean reciprocal rank for person re-identification. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 408–413. IEEE.

Yiquan Wu, Yuhang Liu, Yifei Liu, Ang Li, Siying Zhou, and Kun Kuang. wisdominterrogatory. Available at GitHub.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. *arXiv preprint arXiv:2310.09241*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Nan Xie, Yuelin Bai, Hengyuan Gao, Ziqiang Xue, Feiteng Fang, Qixuan Zhao, Zhijian Li, Liang Zhu, Shiwen Ni, and Min Yang. 2024. Delilaw: A chinese legal

counselling system based on a large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5299–5303.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. Leven: A large-scale chinese legal event detection dataset. *arXiv preprint arXiv:2203.08556*.

Linan Yue, Qi Liu, Yichao Du, Weibo Gao, Ye Liu, and Fangzhou Yao. 2024. Fedjudge: Federated legal large language model. In *International Conference on Database Systems for Advanced Applications*, pages 268–285. Springer.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2024a. A systematic survey of text summarization: From statistical methods to large language models. *arXiv preprint arXiv:2406.11289*.

Liang Zhang, Jionghao Lin, Ziyi Kuang, Sheng Xu, Mohammed Yeasin, and Xiangen Hu. 2024b. Spl: A socratic playground for learning powered by large language mode. *arXiv preprint arXiv:2406.13919*.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

JiDong Ge Zhiwei Fei, Zongwen Shen. 2023. Legal-chatglm: chatglm. https://github.com/NJU-LegalAI/Legal-ChatGLM.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143.

A Statistical data of legal LLMs

Table 1: Statistical data of legal large language models in recent years.

'Release Time' refers to the date when the corresponding model was officially released.
'Publicly Available' indicates that the model code or checkpoints have been open-sourced, while 'Closed Source' means the opposite.
'Open Knowledge Base' refers to the training dataset released with the model.'Hardware' refers to the resources required for the publicly announced model training. In this table, * represents undisclosed content, / indicates no such content, and bold content represents independently constructed and open-sourced datasets.

Model		Release Time	Size (B)	Base Model	Open Knowledge Base	Hardware (GPUs)
Public Available	Lawyer LLaMa (Huang et al., 2023b)	2023/5/24	7/13	LLaMA-7B /Chinese-LLaMA-13B	1.Alpaca-GPT4 52k 2.The instruction fine-tuning data generated by ChatGPT, includes answers to 2k law examination questions and 5k legal consultation response	*
	ChatLaw (Cui et al., 2023a)	2023/6/28	13/33	Ziya-LLaMA-13B /Anima-33B	*	multiple A100
	Legal-ChatGLM (Zhiwei Fei, 2023)	2023/4/22	6	ChatGLM	Instruction-Tuning data from Lawyer LLaMa	4 * 32G V100
	Law GPT_zh [14]	2023/4/22	6	ChatGLM	Use ChatGPT to clean the CrimeKgAssistant dataset and got 52,000 single - turn Q&A pairs. Use ChatGPT to turn Article 9,000 of China's laws and regulations into Q&A for specific scenarios.	4 * RTX3090
	LexiLaw (Haitao Li, 2024)	2023/5/16	6	ChatGLM	Large-scale general domain text dataset BELLE 1.5M; legal question and answer data; Processed legal regulations and legal reference book data; Processed legal document data.	7 * 40G A100
	LaWGPT (Zhou et al., 2024)	2023/5/13	7	Chinese-LLaMA /ChatGLM	*	8 * 32G V100
	JurisLMs[15]	2023/5/15	*	LLaMA	*	*
	HanFei-1.0(He et al., 2023)	2023/5/31	7	BLO-OMz-7B1	A total of 255,000 fine-tuning instruction data.	8 * 40G A100/A800
	Lychee[16]	2023/6/13	10	GLM-10B	*	*
	Wisdom Interrogatory(Wu et al.)	2023/8/21	7	Baichuan-7B	*	
	Fuzi-ming-cha(Deng et al., 2023)	2023/9/6	6	ChatGLM	*	*
	DISC-Law LLM(Yue et al., 2023)	2023/9/26	13	Baichuan-13B-Base	Legal information extraction, judgment prediction, document summarization, legal question answering, and general data, totaling nearly 300,000 entries.	8 * 40G A800

[14] https://github.com/LiuHC0428/LAW_GPT
[15] https://github.com/SEUIAI/JurisLMs
[16] https://github.com/davidpig/lychee_law

Model		Release Time	Size (B)	Base Model	Open Knowledge Base	Hardware (GPUs)
Public Available	SaulLM-7B(Colombo et al., 2024)	2024/3/26	7	Mistral 7B	1.Datasets: Combine both previously available datasets, such as the FreeLaw subset from The Pile and MultiLegal Pile, as well as data scraped from publicly available sources on the Web, resulting in a 30 billion tokens dataset. 2. Leverage a Mistral-7B-instruct to transform legal texts augmented with metadata into coherent conversations.	256*MI250 AMD GPUs
	FedJudge (Yue et al., 2024)	2024/4/10	7	Baichuan-7b	Dataset: - Court:C3VG - Legal consultation/education company: Lawyer LLaMA	2 * Tesla A100 40G GPUs
	KL3M[17]	2024/4/15	170M /1.7B	GPT-NeoX (i.e., GPT-3 architecture)	US: Legal documents from PACER, eCFR, Federal Register, EDGAR, GovInfo, and USPTO patents. EU: EU Official Journal via EUR-Lex. UK: UK legislation from legislation.gov.uk. Germany: German laws from Bundesgesetzblatt.	12 * RTX4090
	InternLM-Law(Fei et al., 2024)	2024/7/21	7	Qwen-1.5-72B	1.Two - stage training pipeline: First, InternLM2 - chat is trained on general and legal tasks. Then, it's further trained on high - quality legal tasks. 2.Legal Data Sources a) Lgeal NLP Data:22 distinct legal-related tasks, yielding a comprehensive dataset consisting of 440K samples. b) Legal Consultation Data: 6 million records obtained from various online platforms. c) Chinese laws & Regulations Data: 100K entries.	64 * 80G A100
	LawLLM (Shu et al., 2024)	2024/7/27	7	Gemma-7B	Dataset:Caselaw Transform training legal cases into high-dimensional vectors using the OpenAI Embedding model. Use a GPT-4 model to extract core information and summarize each case.	*
	AgentCorut (Chen et al., 2024)	2024/8/15	*	*	LLM - driven agent technology in legal scenarios.	*
	DeliLaw [like-framework] (Xie et al., 2024)	2024/8/1	6	ChatGLM2-6b	The user's question is categorized into four types by the classification model. Based on the category, it's sent to the right module. For the LawQuestion category, the Law Retriever gets laws, mixes them with the question, and feeds the combination to the fine - tuned Legal LLM for the final answer.	*

[17] https://huggingface.co/alea-institute/kl3m-002-170m

Model		Release Time	Size (B)	Base Model	Open Knowledge Base	Hardware (GPUs)
Public Available	ADAPT [framework](Deng et al., 2024)	2024/8/6	7B	Qwen2-7B	To guide LLM to gradually deduce the most appropriate charges and law articles step by step, including Ask, Discriminate, and Predict.	*
Closed Source	PowerLaw-GLM[18]	2023/6/28	130	GLM-130B	*	multiple 40G A100
	ALKAFI-LLAMA3 (Qasem et al., 2024)	2024/12/20	1	Llama-3.2-1B-Instruct	1.Language: Palestine 2.CodeBase: Unsloth 3.Dataset Utilize the ChatGPT API and Gemma API to construct a synthetic dataset of question-answer pairs, comprising 1277 text files sourced from the Official Gazette Bureau's website. This dataset contains 243,841 records, totaling approximately 5 million words, and features a vocabulary of 208,835 unique words.	1*RTX3060
	Huayu-Wanxiang[19]	2023/7/10	*	*	*	*
	Deli[20]	2023/9/26	*	*	*	*
	Tiandi[21]	2023/10/12	*	*	*	*
	Tongyi Farui[22]	2023/10/31	*	*	*	*
	ZhiAI[23]	2023/11/18	*	*	*	*
	BAI-Law-13B[24]	2023/12/13	13	Llama2	*	*
	AlphaGPT[25]	2024/2/27	*	*	*	*
	JURU[26]	2024/3/26	*	Sabiá - 2 Small	1.Language: Portuguese 2.Employ the t5x and seqio frameworks for the proposed pretraining. 3.Focus on Brazilian law and general knowledge with Brazilian standardized multiple-choice exams in Portuguese.	A cluster of f TPUs v2-128
	FaXingBao[27]	2024/3	*	*	*	*
	YuanFa ZhiNeng[28]	2024/9	*	Pre-training from scratch	*	*
	FaXin[29]	2024/11/15	*	Pre-training from scratch	*	*
	MetaLaw[30]	*	*	*	*	*
	Shuimu zhifa[31]	2024/8/4	*	*	*	*

[18] https://powerlaw.ai/
[19] https://wanxiang.thunisoft.com/wanxiang/
[20] http://delilegal.com/
[21] https://www.hw99.com/
[22] https://tongyi.aliyun.com/farui
[23] https://www.zhiexa.com/
[24] chen2024alignment
[25] https://www.icourt.cc/prac-tag/357.html
[26] junior2024juru
[27] https://ailegal.baidu.com/m/legalaibot
[28] https://www.law.pku.edu.cn/docs/2023-11/20231120120341542267.pdf
[29] https://ai-bot.cn/faxin-legal-ai-llm/
[30] https://meta.law/
[31] https://www.shuimuzhifa.com/

B AI Governance related Documents

Table 2: AI Governance related Documents

Name	Type	Country/ Region	Date	Authority/ Organization	Risks
Interim Measures for the Administration of Generative Artificial Intelligence Services[32]	Laws and Regulations - Effect	China	2023-08-15	The CAC and other six department	3,2,14
Practice Guidelines for Cybersecurity Standards - Identification Methods for Contents of Generative Artificial Intelligence Services[33]	Standards/ Guidelines	China	2024-08-25	TC260	14
TC260 - 003 Basic Security Requirements for Generative Artificial Intelligence Service[34]	Standards/ Guidelines	China	2024-03-01	TC260	4,3,13,11,17,2
Artificial intelligence-Code of practice for data labeling of machine learning[35]	Standards/ Guidelines	China	2024-12-01	TC28	3
harmonised rules and regulations on artificial intelligence[36]	Laws and Regulations - Effect	EU	2024-06-13	EU	11
Cyber security risks to artificial intelligence[37]	Research Reports	UK	2024-05-15	Department for Science, Innovation &Technology	8,17
OWASP Top 10 for LLM Applications 2025[38]	Research Reports	International Organizations	2024-11-18	OWASP	5,8,17,12,14
Machine learning security principles v2[39]	Research Reports	UK	2024-05-22	NCSC	8,17,9,13
Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile[40]	Standards /Guidelines	USA	2024-07-26	NIST	9,8,17
Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile[41]	Standards /Guidelines	USA	2024-07-26	NIST	11,1,2,4,15
Artificial Intelligence Risk Management Framework (AI RMF 1.0)[42]	Standards /Guidelines	USA	2023-01-26	NIST	4,6,8,12,13
Reducing Risks Posed by Synthetic Content[43]	Standards /Guidelines	USA	2024-11-24	NIST	14,15,3

[32] https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
[33] https://www.tc260.org.cn/front/postDetail.html?id=20230825190345
[34] https://www.tc260.org.cn/front/postDetail.html?id=20240301164054
[35] https://std.samr.gov.cn/gb/search/gbDetailed?id=FC816D04FEB462EBE05397BE0A0AD5FA
[36] https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[37] https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai
[38] https://owasp.org/www-project-top-10-for-large-language-model-applications/
[39] https://www.ncsc.gov.uk/blog-post/machine-learning-security-principles-updated
[40] https://csrc.nist.gov/pubs/sp/800/218/a/final
[41] https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence
[42] https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10
[43] https://www.nist.gov/publications/reducing-risks-posed-synthetic-content-overview-technical-approaches-digital-content

Name	Type	Country/ Region	Date	Authority/ Organization	Risks
LLM AI Cybersecurity & Governance Checklist[44]	Research Reports	International Organizations	2024-02-19	OWASP	2,5
The Bletchley Declaration[45]	Forums and Conferences	Multiple Countries	2024-11-01	AI Safety Summit 2023	9,14
Consensus Statement on Red Lines in Artificial Intelligence[46]	Forums and Conferences	Multiple Countries	2024-03-10	"Beijing AI Security International Dialogue" Forum	9,14,8
Hiroshima Process International Guiding Principles for Advanced AI system[47]	Standards /Guidelines	Multiple Countries	2023-10-30	G7	4,11,14
Guidelines for Secure AI System Development[48]	Standards /Guidelines	Multiple Countries	2023-11-26	USA CISA, UK NCSC, the Australian Signals Directorate's Australian Cyber Security Centre (ASD ACSC), the Canadian Centre for Cyber Security (CCCS), the New Zealand National Cyber Security Centre (NCSC-NZ)	17
Hiroshima Process International Code of Conduct for Advanced AI Systems[49]	Standards /Guidelines	Multiple Countries	2023-10-30	G7	1,2,12,14,8
Blueprint for an AI Bill of Right[50]	Policy Documents	USA	2022-10-04	OSTP (Office of Science and Technology Policy)	7,4,11,2,12,6,13
Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems[51]	Policy Documents	USA	2023-04-25	Consumer Financial Protection Bureau	2,4,7,12,14,6,13
Quality Control Standards for Automated Valuation Models[52]	Standards /Guidelines	USA	2023-06-21	CFPB, OCC, FRB, FDIC, NCUA, and FHFA	4,6,14,13

[44] https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.pdf
[45] https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration
[46] https://idais-beijing.baai.ac.cn/?lang=zh
[47] https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system
[48] https://www.cisa.gov/news-events/alerts/2023/11/26/cisa-and-uk-ncsc-unveil-joint-guidelines-secure-ai-system-development
[49] https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems
[50] https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[51] https://www.ftc.gov/legal-library/browse/cases-proceedings/public-statements
[52] https://www.consumerfinance.gov/rules-policy/final-rules/quality-control-standards-for-automated-valuation-models/

Name	Type	Country/ Region	Date	Authority/ Organization	Risks
Colorado's Consumer Artificial Intelligence Act (SB 24-205)[53]	Laws and Regulations - Effect	USA	2024-05-17	Colorado	4,7,6,5,2,12
Safe and Secure Innovation for Frontier Artificial Intelligence Models Act(CA SB 1047)[54]	Laws and Regulations - Drafted/Not Effect	USA	2024-09-03	California	9,10,8,17,18,5
Digital Content Provenance Standards (CA AB 3211)[55]	Laws and Regulations - Effect	USA	2024-08-23	California	4,14,17,3,6
Artificial Intelligence Security Governance Framework v1[56]	Standards /Guidelines	China	2024-09-09	TC 260	1,2,3,6,4,9,5,16,17,14,12 ,11
Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development[57]	Policy Documents	International Organizations	2024-03-21	UN	4,5,11,12,13,14,6,7,18,16
Guidelines on Securing AI Systems[58]	Standards /Guidelines	Singapore	2024-10-15	CSA	1,2,4,6,7,8,9,12,14,15,16 ,17,18,13
Companion Guide on Securing AI Systems[59]	Standards /Guidelines	Singapore	2024-10-15	CSA	1,2,4,6,7,8,9,12,14,15,16 ,17,18,13
ICO consultation series on generative AI and data protection[60]	Research Reports	UK	2024-09-18	ICO	12,11,13,5,6,7,2
Guidance on AI and data protection[61]	Research Reports	UK	2023-03-15	ICO	4,5,6,7,11,12,14,16,17 ,18,13
Discussion paper on GDPR and LLMs[62]	Research Reports	German	2024-07-15	HmbBfDI	2,12,14,13
Checklist for the use of LLM-based chatbots[63]	Standards /Guidelines	German	2023-11-13	HmbBfDI	2,12,11,18,4,14,13
Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models[64]	Research Reports	EU	2024-10-28	EDPB	2,5,7,8,11,14
The Artificial Intelligence and Data Act (AIDA)[65]	Laws and Regulations - Drafted/Not Effect	Canada	2022-06-16	Minister of Innovation, Science and Industry	2,4,7,8,11,12,14,18
The Artificial Intelligence and Data Act (AIDA) – Companion document[66]	Standards /Guidelines	Canada	2023-03-13	Minister of Innovation, Science and Industry	2,4,7,8,11,12,14,18

[53] https://leg.colorado.gov/bills/sb24-205
[54] https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047
[55] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB3211
[56] https://www.tc260.org.cn/front/postDetail.html?id=20240909102807
[57] https://digitallibrary.un.org/record/4043244/?v=pdf
[58] https://www.csa.gov.sg/Tips-Resource/publications/2024/guidelines-on-securing-ai
[59] https://www.csa.gov.sg/Tips-Resource/publications/2024/guidelines-on-securing-ai
[60] https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations
[61] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence
[62] https://datenschutz-hamburg.de/news/hamburger-thesen-zum-personenbezug-in-large-language-models
[63] https://datenschutz-hamburg.de/news/checkliste-zum-einsatz-llm-basierter-chatbots
[64] https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64
[65] https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading
[66] https://ised-isde.canada.ca/site/innovation-better-canada/en

Name	Type	Country/ Region	Date	Authority/ Organization	Risks
Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems[67]	Standards /Guidelines	Canada	2024-09-01	NAN	4,11,8,14,9,13
Australia's AI Ethics Principles[68]	Standards /Guidelines	Australia	2019-01-01	Department of Industry, Science and Resources	12,7,8,6,13
Safe and Responsible AI in Australia (Discussion paper)[69]	Research Reports	Australia	2023-06-01	Department of Industry, Science and Resources	4,11,13,14,2
Safe and responsible Al in Australia consultation - Australian Government's interim response[70]	Research Reports	Australia	2024-01-17	Department of Industry, Science and Resources	1,4,6,7,11,13,15
Select Committee on Adopting Artificial Intelligence[71]	Laws and Regulations - Effec	Australia	2024-03-26	NAN	1,2,4,7,11,12,14,16,18
National framework for the assurance of artificial intelligence in government[72]	Standards /Guidelines	Australia	2024-06-21	Australian, state and territory governments	4,6,11,17,13
AI Guidelines for Business v1[73]	Standards /Guidelines	Japan	2024-04-19	METI, Ministry of Economy, Trade and Industry	1,2,4,7,11,18,14
Methodology for the Risk and Impact Assessment of Artificial Intelligence Systems from the Point of View of Human Rights, Democracy and the Rule of Law (HUDERIA Methodology)[74]	Standards /Guidelines	EU	2024-09-28	CAI, Committee on Artificial Intelligence	4,6,11,14,13
Open letter to UK online service providers regarding Generative AI and chatbots[75]	Standards /Guidelines	UK	2024-11-08	OFCOM	11,14
Introduction to AI Assurance[76]	Standards /Guidelines	UK	2024-02-12	DSIT, Department for Science, Innovation and Technology	4,7,11,12,2
Guidance for using the AI Management Essentials tool[77]	Standards /Guidelines	UK	2024-11-06	DSIT	4,7,5,12,11,13

[67] https://ised-isde.canada.ca/site/ised/en
[68] https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles
[69] https://consult.industry.gov.au/supporting-responsible-ai
[70] https://consult.industry.gov.au/supporting-responsible-ai
[71] https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Adopting_Artificial_Intelligence_AI/AdoptingAI
[72] https://www.finance.gov.au/sites/default/files/2024-06/National-framework-for-the-assurance-of-AI-in-government.pdf
[73] https://www.meti.go.jp/english/press/2024/0419_002.html
[74] https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems
[75] https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content
[76] https://www.gov.uk/government/publications/introduction-to-ai-assurance
[77] https://www.gov.uk/government/consultations/ai-management-essentials-tool

Name	Type	Country/ Region	Date	Authority/ Organization	Risks
Compliance of products with embedded artificial intelligence[78]	Standards /Guidelines	International Organizations	2024-11-04	UNECE	4,8,11,17,13
Model AI Governance Framework for Generative AI[79]	Standards /Guidelines	Singapore	2024-05-24	IMDA	1,2,4,6,9,11,13,17
AI and cyber security: what you need to know[80]	Research Reports	UK	2024-02-13	NCSC	4,11,13,14
Machine learning principles[81]	Standards /Guidelines	UK	2024-05-22	NCSC	4,8,9,11,12,17,18
Artificial Intelligence Risk Governance Report (2024) - Constructing a Practical Scheme for Artificial Intelligence Security Governance Oriented to the Industry[82]	Research Reports	China	2024-12-24	CAICT	3,4,5,6,1,11,12,15,17,18
Global Artificial Intelligence Governance Research Report[83]	Research Reports	China	2024-11-22	World Internet Conference	15,12,11,8,6,1,2,5,17
AB-2013 Generative artificial intelligence: training data transparency[84]	Laws and Regulations - Drafted/Not Effect	UK	2024-09-28	California	1,2,4,5
SB-942 California AI Transparency Act[85]	Laws and Regulations - Drafted/Not Effect	UK	2024-09-19	California	14,12,9,2,5
Ensuring Likeness, Voice, and Image Security Act[86]	Laws and Regulations - Effect	UK	2024-07-01	Tennessee	14,1,2,12
Bill on the use of Artificial Intelligence[87]	Laws and Regulations - Drafted/Not Effect	Brazil	2023-05-03	Senado	4,10,11,14,1,2,5,9,12,15
Fundamental Law on the Development of Artificial Intelligence and the Establishment of Trust[88]	Laws and Regulations - Effect	Korea	2024-12-26	NAK	8,9,10,11,14,1,2,5,12,15 ,18
First Draft-GeneralPurpose AI Code of Practice[89]	Laws and Regulations - Drafted/Not Effect	EU	2024-11-14	European commission	5,11,14,8,9,10,1,2
Fundamental Law on Artificial Intelligence[90]	Laws and Regulations - Drafted/Not Effect	China	2024-07-15	TAIWAN, CHINA	5,2,8,9,11,12,14,7,15

[78] https://unece.org/trade/publications/ece_trade_486
[79] https://aiverifyfoundation.sg/resources/mgf-gen-ai/
[80] https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you-need-to-know
[81] https://www.ncsc.gov.uk/collection/machine-learning-principles
[82] https://www.caict.ac.cn/kxyj/qwfb/ztbg/202412/t20241225_648969.htm
[83] https://cn.wicinternet.org/2024-11/22/content_37693427.htm
[84] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013&_bhlid=9ee694300fd508d6947b3df85a730674cfc893b1
[85] https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942&firstNav=tracking
[86] http://www.capitol.tn.gov
[86] https://www25.senado.leg.br/web/atividade/materias/-/materia/157233
[88] https://www.gov.kr/portal/ntnadmNews/4124183
[89] https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts
[90] https://join.gov.tw/policies/detail/4c714d85-ab9f-4b17-8335-f13b31148dc4

C Awesome-LegalAI-Resources

Table 3: Awesome-LegalAI-Resources

Type		Name	Description	language	Country
General Corpus		MultiLegalPile (Niklaus et al., 2023b)	A 689GB corpus in 24 languages from 17 jurisdictions	multilingual	multinational
		MC4_legal[91]	This dataset contains large text resources (106GB in total) from mc4 filtered for legal data that can be used for pretraining language models	multilingual	multinational
		EurlexResources[92]	This dataset contains large text resources (179GB in total) from EURLEX that can be used for pretraining language models	multilingual	multinational
		LeXFile (Chalkidis et al., 2023)	We created a new, diverse English multinational legal corpus, LeXFiles. It has 11 sub - corpora, covering legislation and case law from 6 major English - speaking legal systems (EU, CoE, Canada, US, UK, India) and contains about 19 billion tokens.	English	multinational
		Pile of Law (Henderson et al., 2022)	A 256GB (and growing) dataset of open-source English-language legal and administrative data, covering court opinions, contracts, administrative rules, and legislative records.	English	Unknown
		Spanish Legal Domain Corpora (Gutiérrez-Fandiño et al., 2021)	Our corpora comprises multiple digital resources and it has a total of 8.9GB of textual data.	Spanish	Spanish
		GeLeCo[93]	A large German legal corpus for research, teaching and translation. It comprises federal laws, administrative regulations and court decisions from three online databases of the German Federal Ministry of Justice and Consumer Protection and the Federal Office of Justice.	German	German
		CourtListener[94]	The original Court Listener dataset contains all court opinions from US courts, covering 406 out of 423 jurisdictions from 1754 to the present. It's updated regularly with new opinions and digitized archives.	English	America
Evaluation Benchmark	Multi Legal Task	LegalLAMA (Chalkidis et al., 2023)	A diverse probing benchmark suite comprising 8 sub-tasks that aims to assess the acquaintance of legal knowledge that PLMs acquired in pretraining.	English	multinational
		LexGLUE (Chalkidis et al., 2021)	LexGLUE comprises seven datasets: ECtHR Task A and B, SCOTUS, EUR-LEX, LEDGAR, UNFAIR-ToS, and CaseHOLD that are available for re-use and re-share with appropriate attribution.	English	multinational
		LEXTREME (Niklaus et al., 2023a)	The dataset consists of 11 diverse multilingual legal NLU datasets. 6 datasets have one single configuration and 5 datasets have two or three configurations. This leads to a total of 18 tasks.	multilingual	multinational

[91] https://huggingface.co/datasets/joelito/legal-mc4
[92] https://huggingface.co/datasets/joelito/eurlex_resources
[93] https://github.com/antcont/GeLeCo
[94] https://www.courtlistener.com/help/api/bulk-data/

Type		Name	Description	language	Country
Evaluation Benchmark	Multi Legal Task	LegalBench (Guha et al., 2022)	A collaborative benchmark intended to evaluate English large language models on legal reasoning and legal text-based tasks. LegalBench currently consists of more than 90 tasks.	English	multinational
		LBOX OPEN (Hwang et al., 2022)	This paper presents the first large - scale Korean legal AI dataset benchmark, LBOX OPEN. It includes one legal corpus, two classification tasks, two legal judgement prediction (LJP) tasks, and one summarization task.	Korean	Korean
		GENTLE (Aoyama et al., 2023)	We present GENTLE, a new mixed-genre English challenge corpus totaling 17K tokens and consisting of 8 unusual text types for out-of domain evaluation: dictionary entries, esports commentaries, legal documents, medical notes, poetry, mathematical proofs, syllabuses, and threat letters.	English	Unknown
		SCALE (Rasiah et al., 2023)	In this paper, we introduce a novel NLP benchmark challenging current LLMs in four aspects: handling long documents (up to 50K tokens), applying domain - specific legal knowledge, achieving multilingual understanding (five languages), and multitasking (including legal document IR, court view generation, decision summarization, citation extraction, and eight text classification tasks).	multilingual	Switzerland
	Legal Case Retrieval	LeCaRD (Ma et al., 2021)	LeCaRD composes of 107 query cases and 10,700 candidate cases selected from a corpus of over 43,000 Chinese criminal judgements.	Chinese	China
		LeCaRDv2[95]	LeCaRDv2 is one of the largest Chinese legal case retrieval datasets with the widest coverage of criminal charges. The dataset comprises 800 query cases and 55,192 candidate cases extracted from 4.3 million criminal case documents.	Chinese	China
		COLIEE (Kim et al., 2022)	The Competition on Legal Information Extraction/Entailment (COLIEE) is an annual international competition whose aim is to achieve state-of-the-art methods for legal text processing. Task 1 is the legal case retrieval task. Task 3 is the statute law retrieval task.	English /Japanese	Canada /Japan
		document-similarity (Bhattacharya et al., 2022)	The task here is to calculate a similarity score (in the range 0-1) between two case documents. The dataset collected 53, 210 publicly available case documents from the Supreme Court of India and and 12, 814 Acts from the Indian judiciary.	English	India
	Question Answering	JEC-QA (Zhong et al., 2020)	the largest question answering dataset in the legal domain, collected from the National Judicial Examination of China. There are 26,365 questions in JEC-QA.	Chinese	China
		CaseHOLD (Zheng et al., 2021)	This CaseHOLD dataset provides 53,000+ multiple choice questions with prompts from a judicial decision and multiple potential holdings, one of which is correct, that could be cited.	English	America
		SARA (Holzenberger et al., 2020)	A novel dataset based on US tax law, together with test cases.	English	America
		PrivacyQA (Ravichander et al., 2019)	PrivacyQA is a corpus consisting of 1750 questions about the contents of privacy policies, paired with expert annotations.	English	America

[95] https://github.com/THUIR/LeCaRDv2

Type		Name	Description	language	Country
Evaluation Benchmark	Legal Case Entailment	COLIEE (Kim et al., 2022)	The Competition on Legal Information Extraction/Entailment (COLIEE) is an annual international competition whose aim is to achieve state-of-the-art methods for legal text processing. Task 2 is the legal case entailment task. Task 4 is the legal textual entailment data corpus.	English /Japanese	Canada /Japan
		Legal Linking (Shaffer and Mayhew, 2019)	This paper describes a dataset and baseline systems for linking paragraphs from court cases to clauses or amendments in the US Constitution.	English	America
	Document Classification	CAIL2018 (Xiao et al., 2018)	CAIL2018 contains more than 2.6 million criminal cases published by the Supreme People's Court of China. It consists of applicable law articles, charges, and prison terms, which are expected to be inferred according to the fact descriptions of cases.	Chinese	China
		ECHR (Chalkidis et al., 2019a)	This paper contributes a new publicly available English legal judgment prediction dataset of cases from the European Court of Human Rights (11.5k cases).	English	European
		Swiss-Judgment-Prediction (Niklaus et al., 2021)	The paper publicly release a multilingual (German, French, and Italian), diachronic (2000-2020) corpus of 85K cases from the Federal Supreme Court of Switzer- land (FSCS).	multilingual	multinational
		German Legal Decision Corpora (Urchs et al., 2021)	To meet this need for publicly available German legal text corpora this paper presents two German legal text corpora. The first corpus contains 32,748 decisions from 131 German courts, enriched with metadata. The second one is a subset of the first corpus and consists of 200 randomly chosen judgements.	German	German
		EURLEX57K (Chalkidis et al., 2019b)	We release a new dataset of 57k legislative documents from EUR-LEX, the European Union's public document database, annotated with concepts from EUROVOC, a multidisciplinary thesaurus.	English	European
		German rental agreements (Glaser et al., 2018)	601 sentences from the tenancy law of the German Civil Code and 312 sentences, classified according to a semantic type system consisting of 9 different classes, from German rental agreements.	English	German
	Summarization	BillSum (Kornilova and Eidelman, 2019)	We introduce the BillSum dataset, which contains a primary corpus of 22,218 US Congressional bills and reference summaries split into a train and a test set.	English	America
		EUR-Lex-Sum (Aumiller et al., 2022)	We obtain up to 1,500 document/summary pairs per language, including a subset of 375 crosslingually aligned legal acts with texts available in all 24 languages.	multilingual	European
		Plain English Summarization of Contracts (Manor and Li, 2019)	The dataset we propose contains 446 sets of parallel text.	English	American
		Summarization-of-Privacy-Policies (Keymanesh et al., 2020)	This dataset was extracted from the text of privacy policy, terms of service, and cookie policy of 151 companies. The Points and Plain English Summaries are extracted from tosdr.org.	English	Unknown
		Multi-LexSum (Shen et al., 2022)	We introduce Multi-LexSum, a collection of 9,280 expert-authored summaries drawn from ongoing CRLC writing.	English	Unknown

Type		Name	Description	language	Country
Evaluation Benchmark	Entity extraction	CDJUR-BR (Mauricio et al., 2023)	We describe the development of the Golden Collection of the Brazilian Judiciary (CDJUR-BR) contemplating a set of fine-grained named entities that have been annotated by experts in legal documents. This contains 44,526 annotations for 21 entities.	Portuguese	Brazilian
		Extracting Contract Elements (Chalkidis et al., 2017)	The paper describes and is accompanied by a new benchmark dataset of approximately 3,500 English contracts with gold contract element annotations.	English	England
		LEVEN (Yao et al., 2022)	LEVEN contains 108 event types in total, including 64 charge-oriented events and 44 general events. Their distribution is shown below.	Chinese	China
	Others	MAUD (Wang et al., 2023b)	To address this challenge, we introduce the Merger Agreement Understanding Dataset (MAUD), an expert-annotated reading comprehension dataset based on the American Bar Association's 2021 Public Target Deal Points Study, with over 39,000 examples and over 47,000 total annotations.	English	America
		VerbCL (Rossi et al., 2021)	This paper presents a new dataset that consists of the citation graph of court opinions, which cite previously published court opinions in support of their arguments.	English	America
		MultiLegalSBD (Brugger et al., 2023)	Sentence Boundary Detection (SBD) is one of the foundational building blocks of Natural Language Processing (NLP), with incorrectly split sentences heavily influencing the output quality of downstream tasks. We curated a diverse multilingual legal dataset consisting of over 130'000 annotated sentences in 6 languages.	multilingual	multinational
		FairLex (Chalkidis et al., 2022)	Our benchmarks cover four jurisdictions (European Council, USA, Switzerland, and China), five languages (English, German, French, Italian and Chinese) and fairness across five attributes (gender, age, region, language, and legal area).	multilingual	multinational
		ContractNLI (Koreeda and Manning, 2021)	In this work, we propose documentlevel natural language inference (NLI) for contracts, a novel, real-world application of NLI that addresses such problems. We annotated and release the largest corpus to date consisting of 607 annotated contracts.	English	America
		Demosthen (Grundler et al., 2022)	A novel corpus for argument mining in legal documents, composed of 40 decisions of the Court of Justice of the European Union on matters of fiscal state ai.	English	European

D Details of the 18 LLM Risks

Table 4: Details of the 18 LLM Risks

Type	NO.	Description	Example	Sources
Trainning Data	1	When copyrighted text, images, or other media are used in the training dataset without permission, models risk infringing on intellectual property rights.	GitHub Copilot was criticized for occasionally generating code that was identical or extremely similar to existing copyrighted code, sparking debates on fair use.	Developers warned: GitHub Copilot code may be licensed (TechTarget, 2022)[96]
	2	Training data might contain personal data (e.g., names, addresses, sexual orientation), breaching privacy regulations like GDPR.	Italy's data protection authority briefly banned ChatGPT in 2023 over concerns that personal information was being collected without adequate legal basis.	Report of the work undertaken by the ChatGPT Taskforce (EDPB, 2023)[97]
	3	The annotation process can expose human annotators to confidential or sensitive content (e.g., trade secrets, personal data). Annotators might also introduce bias or errors.	Annotation practices have been found to encode gender biases into AI systems. For instance, sentiment analysis models have been highlighted for biased results, where sentiments expressed by marginalized groups are labeled more negatively.	The Forgotten Layers: How Hidden AI Biases Are Lurking in Dataset Annotation Practices (UNITE, 2024)[98]
	4	Inappropriate content refers to content in the training data that may be controversial, illegal, or harmful. The risks of inappropriate content stem from various sources, including but not limited to the issues outlined in Risks 1–3, as well as potential causes such as data poisoning or unreasonable data scraping practices.	Researchers from New York University published a study in *Nature Medicine*, stating that replacing just 0.001% of the training tokens with incorrect information can make the trained model more likely to spread false medical content.	Medical large language models are vulnerable to data-poisoning attacks(nature, 2025)[99]
	5	Large training sets and their accompanying metadata can be stolen or leaked-especially if cloud storage or annotation platforms are compromised. Please note that the risk mentioned here is different from the data breach in Risk 12. The latter primarily refers to the model's output containing leaked personal information, protected works of others, trade secrets, and similar content.	Wiz Research discovered that when Microsoft's AI researchers were releasing open - source training data on GitHub, due to a misconfigured SAS token, 38 terabytes of private data were accidentally exposed. There are also security risks, and relevant security recommendations were put forward.	38TB of data accidentally exposed by Microsoft AI researchers(WIZ, 2023)[100]
The LLM Itself	6	Neural networks-especially large ones-are often "black boxes," making it difficult to explain how they arrive at specific outputs.	the Black Box Problem in AI poses significant challenges for cybersecurity by creating issues around trust, accountability, ethics, debugging, compliance, and vulnerability to data poisoning.	Navigating the AI Black Box Problem(Gibraltar, 2024)[101]
	7	Discriminatory of LLMs mainly stems from risks at the training data level. In addition, factors such as model design and the training process may also lead to the discriminatory.	Microsoft's Tay chatbot quickly started to generate racist content on Twitter after interacting with trolls, demonstrating how biases can be learned and repeated.	Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day(The Verge, 2016)[102]

[96] https://www.techtarget.com/searchsoftwarequality/news/252526359/Developers-warned-GitHub-Copilot-code-may-be-licensed
[97] https://www.edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-chatgpt-taskforce_en
[98] https://www.unite.ai/the-forgotten-layers-how-hidden-ai-biases-are-lurking-in-dataset-annotation-practices/?utm_source=chatgpt.com
[99] https://www.nature.com/articles/s41591-024-03445-1
[100] https://www.wiz.io/blog/38-terabytes-of-private-data-accidentally-exposed-by-microsoft-ai-researchers
[101] https://gibraltarsolutions.com/blog/navigating-the-ai-black-box-problem/
[102] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Type	NO.	Description	Example	Sources
The LLM Itself	8	Small changes or "adversarial" prompts can cause drastic shifts in outputs, meaning the model can fail badly outside typical scenarios.	The research found that even the slightest perturbation of the prompt, such as adding a space at the end of the prompt, may cause the large language model to change its answer.	How Small Changes and Jailbreaks Affect Large Language Model Performance(Abel et al., 2024)[103]
	9	Attackers might manipulate model parameters ("model poisoning") or training data to insert backdoors or alter behavior.	Research has demonstrated that by injecting certain trigger phrases during training, an LLM can be forced to produce malicious outputs when prompted with that trigger.	Hidden Backdoors in Neural Networks (Gu et al., 2017)[104]
	10	Errors, biases, or vulnerabilities that appear at one stage of development can propagate and amplify in subsequent versions of a model (or downstream tasks).	If a poorly vetted LLM is used to create training data for the next generation of models, the original issues can become more deeply ingrained.	On the Dangers of Stochastic Parrots (Bender et al., 2021)[105]
Generated Cotent	11	LLMs may generate hate speech, radicalization content, or instructions for violence if prompted or misused.	The report stating that large language models, including OpenAI's GPT - 3.5, GPT - 4, and Meta's Llama 2, show obvious biases against women in their generated content.	Large language models generate biased content, warn researchers(TechXplore, 2024)[106]
	12	The LLM might inadvertently reveal private or proprietary information it was trained on or that was input by other users.	Samsung employees reportedly entered sensitive source code into ChatGPT, which then became part of the service's broader training data.	Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak (Bloomberg, 2023)[107]
	13	LLMs sometimes generate plausible-sounding but factually incorrect answers.	OpenAI's audio transcription tool, Whisper, experiences "hallucinations" in high - risk situations, generating content that is not present in its training materials.	OpenAI's Whisper AI 'hallucinates' in high-risk situations(tom's guide, 2024)[108]
	14	Malicious actors can use LLMs for phishing emails, social engineering, generating malware, or other harmful activities.	Security researchers have shown that ChatGPT-like models can assist in writing malicious code or highly personalized phishing messages.	ChatGPT tool could be abused by scammers and hackers (BBC, 2023)[109]
	15	Overreliance on LLMs outputs (e.g., in journalism or healthcare) might lead to skill degradation, job displacement, or mental health concerns if LLMs replace human interaction.	Some mental health app chatbots rely heavily on LLMs for conversations, raising concerns about quality of care and accountability.	Chatbot therapy is risky. It's also not useless. (Vox, 2023)[110]
Others	16	Training and running LLMs is resource-intensive, leading to high operational costs, potential single-vendor reliance, and large carbon footprints.	The training of GPT-4 was estimated to produce significant CO emissions, raising sustainability concerns.	Reconciling the contrasting narratives on the environmental impact of large language models (Ren et al., 2024)[111]
	17	Vulnerabilities in hardware (GPUs, chips) or software libraries used to develop LLMs could compromise the entire pipeline. Shortages or geopolitical issues can disrupt hardware supply.	Content related to LLM supply chain attacks, including real cases such as OpenAI's Python library vulnerability, the abuse of the PyPI code library dependency chain, the ChatGPT plugin vulnerability, and the cracking of Hugging Face's safetensors.	LLM Supply Chain Attack: Prevention Strategies(Coblat, 2024)[112]
	18	LLMs are often deployed as part of larger systems-failure in one component (e.g., cloud servers, network infrastructure) can bring down critical services.	OWASP has released the top ten vulnerabilities of LLM applications, including prompt injection, insufficient sandboxing, unauthorized code execution, server - side request forgery, over - reliance on LLM - generated content, and insufficient AI alignment, etc.	The Dangers of AI: OWASP Releases Top 10 Vulnerabilities for LLM Applications(PantaSecurity, 2023)[113]

[103] https://arxiv.org/html/2401.03729v2
[104] https://arxiv.org/abs/1708.06733
[105] https://dl.acm.org/doi/10.1145/3442188.3445922
[106] https://techxplore.com/news/2024-04-large-language-generate-biased-content.amp
[107] https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak
[108] https://www.tomsguide.com/ai/openais-whisper-model-is-reportedly-hallucinating-in-high-risk-situations
[109] https://www.bbc.com/news/technology-67614065
[110] https://www.vox.com/technology/2023/12/14/24000435/chatbot-therapy-risks-and-potential
[111] https://www.nature.com/articles/s41598-024-76682-6
[112] https://www.cobalt.io/blog/llm-supply-chain-attack-prevention-strategies
[113] https://www.pentasecurity.com/blog/dangers-ai-owasp-top-10-llm/

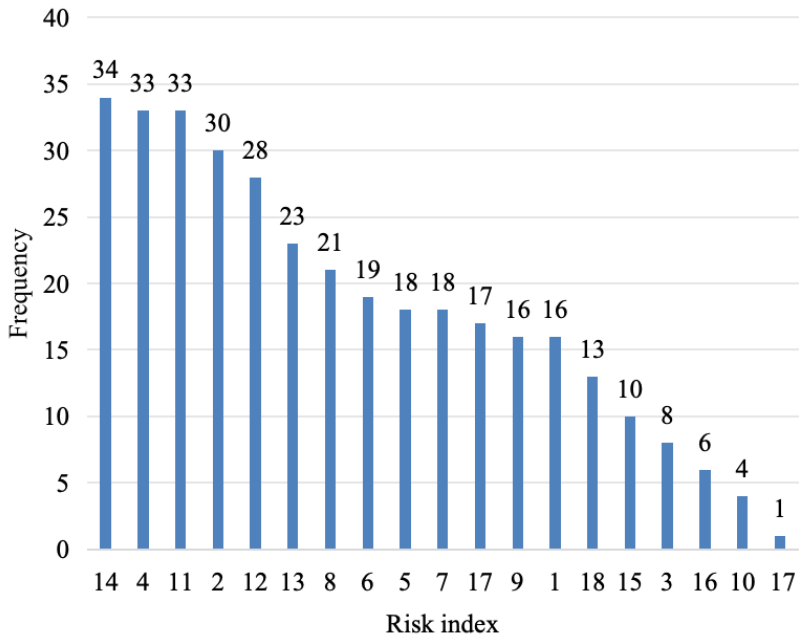
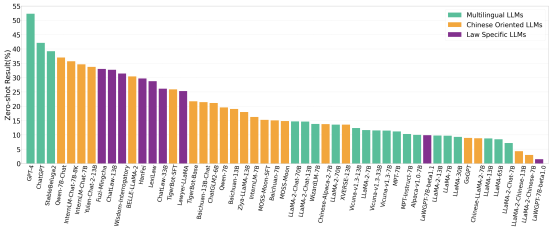
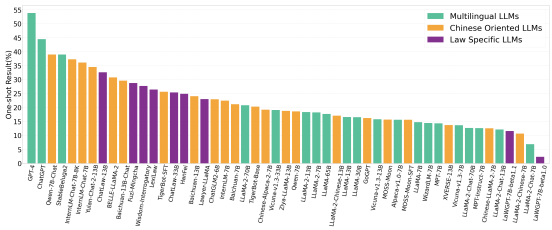


Figure 3: The Frequency of LLM Risks in Documents

E Benchmark Results



(a). Average performance (zero-shot) of 51 LLMs evaluated on LawBench.



(b). Average performance (one-shot) of 51 LLMs evaluated on LawBench.

Figure 4: Model Performance in LawBench.

Table 5: Average performance for each LLM over the different LegalBench categories. The first block of rows corresponds to large commercial models, the second block corresponds to models in the 11B-13B range, the third block corresponds to models in the 6B-7B range, and the final block corresponds to models in the 2B-3B range. The columns correspond (in order) to: issue-spotting, rule-recall, rule-conclusion, interpretation, and rhetorical-understanding. For each class of models (large, 13B, 7B, and 3B).

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical
GPT-4	82.9	59.2	89.9	75.2	79.4
GPT-3.5	60.9	46.3	78.0	72.6	66.7
Claude-1	58.1	57.7	79.5	67.4	68.9
Flan-T5-XXL	66.0	36.0	63.3	64.4	70.7
LLaMA-2-13B	50.2	37.7	59.3	50.9	54.9
OPT-13B	52.9	28.4	45.0	45.1	43.2
Vicuna-13B-16k	34.3	29.4	34.9	40.0	30.1
WizardLM-13B	24.1	38.0	62.6	50.9	59.8
BLOOM-7B	50.6	24.1	47.2	42.8	40.7
Falcon-7B-Instruct	51.3	25.0	52.9	46.3	44.2
Incite-7B-Base	50.1	36.2	47.0	46.6	40.9
Incite-7B-Instruct	54.9	35.6	52.9	54.5	45.1
LLaMA-2-7B	50.2	33.7	55.9	47.7	47.7
MPT-7B-8k-Instruct	54.3	25.9	48.9	42.1	44.3
OPT-6.7B	52.4	23.1	46.3	48.9	42.2
Vicuna-7B-16k	3.9	14.0	35.6	28.1	14.0
BLOOM-3B	47.4	20.6	45.0	45.0	36.4
Flan-T5-XL	56.8	31.7	52.1	51.4	67.4
Incite-3B-Instruct	51.1	26.9	47.4	49.6	40.2
OPT-2.7B	53.7	22.2	46.0	44.4	39.8

Table 6: Zero-shot performance(%) of various models at Memorization, Understanding, and Logic Inference level. Best preformance in each column is marked bold.

Model	Memorization(Acc.)			Understanding(Acc.)					Logic Inference(Acc.)					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
GPT-4	**34.0**	35.4	14.0	79.8	**51.0**	**94.0**	78.0	**96.2**	**80.3**	68.3	**53.7**	**33.2**	**66.0**	**57.8**
Qwen-14B-Chat	28.0	38.6	11.4	**93.4**	45.3	90.0	**85.6**	91.8	80.2	**91.0**	27.9	31.6	44.7	50.4
Qwen-7B-Chat	22.8	**38.9**	8.4	79.8	43.3	87.0	67.2	92.0	79.2	83.9	53.2	24.2	36.3	45.0
ChatGPT	19.0	25.6	9.0	56.8	42.3	87.0	76.0	82.2	77.7	60.3	23.0	19.4	39.6	38.2
InternLM-7B-Chat	20.4	35.4	11.0	61.4	42.3	89.0	49.4	53.8	79.3	77.9	28.8	23.8	38.3	30.0
Baichuan-13B-Chat	14.6	33.9	10.0	54.2	35.0	72.0	62.2	75.4	77.0	58.0	41.8	20.2	33.5	21.0
ChatGLM3	19.2	28.9	7.7	41.0	34.3	80.0	62.8	81.4	73.4	61.2	19.4	21.4	25.6	37.0
Baichuan-13B-base	22.6	23.0	9.0	43.2	26.7	75.0	59.2	74.4	58.3	25.6	12.5	23.8	31.0	19.6
Fuzi-Mingcha	13.0	25.0	6.7	62.0	29.0	61.0	46.4	24.8	68.0	58.6	25.5	16.0	28.9	20.4
ChatLaw-33B	16.0	25.9	7.0	51.4	32.3	76.0	67.6	62.0	60.6	32.9	23.0	15.4	23.6	37.6
ChatGLM2	28.2	13.6	16.4	22.4	24.0	61.0	40.0	29.8	77.2	54.4	24.8	19.8	27.7	8.6
Chinese-Alpaca-2-7B	19.8	24.8	**19.7**	25.0	33.3	61.0	46.6	24.2	66.8	39.4	20.6	16.4	18.0	26.6
BELLE-LLAMA-2-Chat	15.0	25.7	7.0	31.4	27.3	77.0	61.6	46.2	64.1	47.3	8.2	19.8	33.2	24.4
XVERSE-13B	25.4	29.0	12.0	47.0	21.7	71.0	48.2	32.4	54.9	44.7	9.9	19.2	27.7	14.6
TigerBot-base	16.6	27.5	9.0	22.4	27.0	58.0	57.0	24.6	71.5	35.7	18.3	19.0	31.2	18.8

Table 7: Zero-shot performance(%) of various models at Discrimination, Generation, and Ethic level. Best preformance in each column is marked bold.

Model	Discrimination(Acc.)		Generation(Rough-L)				Ethic(Acc.)			Average	Rank
	4-1	4-2	5-1	5-2	5-3	5-4	6-1	6-2	6-3		
GPT-4	35.8	**39.1**	25.0	16.0	**38.3**	13.6	**65.2**	**55.2**	**75.8**	**52.4**	**1**
Qwen-14B-Chat	30.0	31.9	33.9	23.1	36.0	**19.1**	29.2	42.0	63.0	48.6	2
Qwen-7B-Chat	21.0	28.6	30.8	19.0	34.7	18.3	22.1	38.9	56.8	44.8	3
ChatGPT	28.4	22.0	22.8	13.1	34.3	13.1	33.7	32.1	55.8	39.6	4
InternLM-7B-Chat	**37.0**	9.9	19.6	2.6	29.2	11.8	22.7	27.8	47.4	36.9	5
Baichuan-13B-Chat	24.4	20.4	29.2	24.2	35.7	16.0	16.4	22.0	40.8	36.4	6
ChatGLM3	25.2	14.1	28.3	17.0	29.7	14.4	21.2	29.6	49.6	35.8	7
Baichuan-13B-base	15.6	23.0	21.5	**27.8**	24.0	11.8	17.3	28.6	47.0	31.3	8
Fuzi-Mingcha	20.0	16.1	**57.8**	**27.8**	21.4	17.3	10.8	13.1	25.0	30.2	9
ChatLaw-33B	10.0	17.1	23.8	9.9	15.2	13.3	15.3	19.1	34.2	30.0	10
ChatGLM2	20.2	21.1	28.4	15.5	24.1	14.0	36.8	27.2	52.2	29.9	11
Chinese-Alpaca-2-7B	27.8	24.7	28.6	15.7	31.2	14.6	21.5	28.4	40.4	29.4	12
BELLE-LLAMA-2-Chat	3.6	20.4	28.0	11.4	25.4	15.3	13.8	16.6	30.4	28.4	13
XVERSE-13B	10.4	12.2	12.1	13.9	6.8	19.0	19.9	29.4	55.0	27.7	14
TigerBot-base	25.8	23.0	20.8	11.3	34.5	12.6	16.3	19.0	39.2	27.3	15

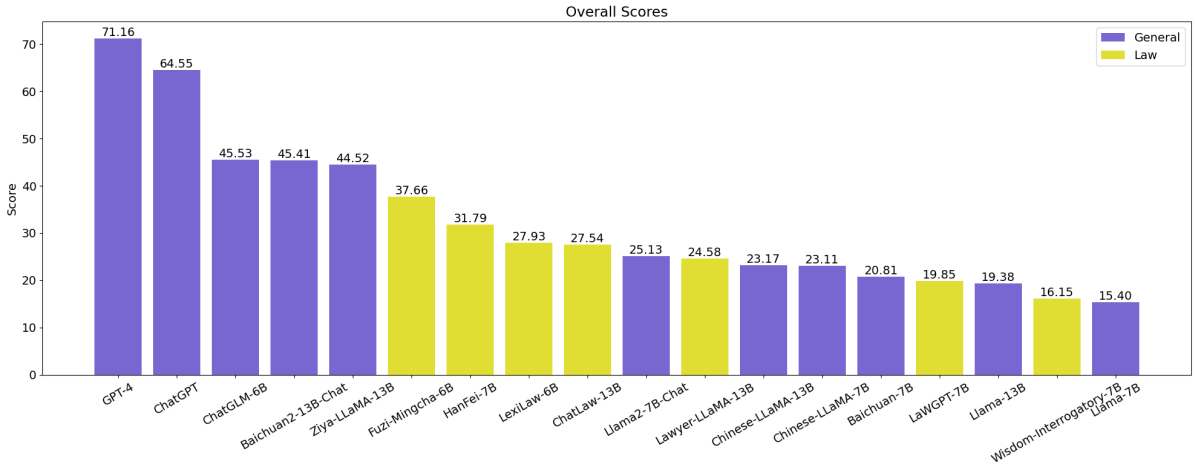


Figure 5: Model Performance in LaiW.

F Evaluation Metrics

Table 8: Summary of Evaluation Metrics

Metric	Description	Formula
Recall	Measures the proportion of actual positive cases correctly identified.	Recall $= \frac{TP}{TP+FN}$
Accuracy	Measures the proportion of correctly predicted samples out of all samples.	Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
F1-Score	Harmonic mean of precision and recall.	$F1-Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$
MAE	Measures the average absolute difference between predictions and actual values.	MAE $= \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
NDCG@k	Normalized Discounted Cumulative Gain, evaluates ranking quality.	NDCG@k $= \frac{DCG@k}{IDCG@k}$, $DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$
MRR	Measures the inverse rank of the first relevant result.	MRR $= \frac{1}{\lvert Q \rvert}\sum_{i=1}^{\lvert Q \rvert} \frac{1}{\text{rank}_i}$

G Legal LLM Training Process

As shown in Figure 6, most legal LLMs are built on base models and are trained using legal datasets following the "pre-training, fine-tuning, and retrieval enhancement" process. For legal LLMs, the pre-training process aids in understanding general language, the fine-tuning process helps to comprehend legal language and grasp the legal logic within it, and the RAG (Retrieval-Augmented Generation) contributes to the model's ability to provide answers based on the precise legal knowledge it has retrieved.

Continual Pre-training Continual Pre-training refers to the process of further training the model on a large-scale unlabelled dataset based on the pre-trained base model, with the aim of enhancing the model's performance in the comprehension of legal texts. This stage mainly employs self-supervised or unsupervised learning methods, enabling the model to automatically learn more meaningful feature representations from a larger dataset.

Continual pre-training primarily addresses the data discrepancy between the pre-training corpus and the fine-tuning corpus for downstream tasks. The continual pre-training of legal LLMs typically serves two purposes:

1. Improve the base model's understanding of different native languages. By introducing various general-domain datasets in specific native languages (such as BELLE (Ji et al., 2023a), alpaca _chinese(Cui et al., 2023c), etc.) to the multilingual or English base models, the model could learn to comprehend and generate texts in languages other than English. Alternatively, researchers may use continual training checkpoints and pretrained LLMs in another language, such as Chinese-LLaMA (Cui et al., 2023b), ChatGLM (GLM et al., 2024), and Baichuan (Yang et al., 2023), to save time in training the model to understand another language.

2. Enhance the base model's understanding of the legal field. Since pre-trained language models use massive heterogeneous corpora, it is common to continue pre-training on a large amount of unlabelled, domain-specific legal text to expand the legal vocabulary. This step is known as Domain-Adaptive Pretraining (DAPT). Research shows that DAPT can improve the performance of downstream tasks, especially when the distribution of pre-training corpus and domain corpus is larger (Gururangan et al., 2020).

Fine-tuning Fine-tuning refers to the process of making precise adjustments to the pre-trained model using a small-scale labeled dataset, enabling the model to adapt better to specific tasks (Min et al., 2023). This stage is primarily achieved through supervised learning, allowing the model to learn task-specific feature representations.

In the fine-tuning of legal LLMs, researchers typically direct the existing general-domain LLMs to perform supervised fine-tuning (SFT) on legal datasets, such as legal documents, legal articles, and high-quality legal question-and-answer datasets (Cui et al., 2023a; Haitao Li, 2024; Huang et al., 2023b). These datasets are labeled by law professionals. This method allows the general-domain LLMs to learn rich legal domain knowledge and aligns the model's understanding of semantics with the legal field. Using solely this method without DAPT can reduce the cost of training legal LLMs, but its efficacy in the legal field may not be as good as those of the pre-trained legal LLMs with fine-tuning.

Retrieval-Augmentation General domain LLMs often suffer from hallucinations (Ji et al., 2023b; Huang et al., 2023a), where they often generate statements that either do not adhere to the original text or fail to align with facts. In the legal field, the content generated by LLMs needs to be highly knowledgeable and reliable. However, the hallucination issue can lead to the creation of fabricated legal provisions or falsifyied legal facts, rendering LLMs unreliable (Magesh et al., 2024). At the same time, legal knowledge is continuously updated. If we simply rely on pre-training and fine-tuning LLMs, the models may remember and keep forever the outdated knowledge learned in the corpus used for training, resulting in knowledge conflicts and an inability to quickly learn updated domain knowledge. To address these issues, many researchers have optimized legal LLMs using a retrieval-augmented approach (Cui et al., 2023a; Huang et al., 2023b; Cui et al., 2024).

Retrieval-augmentation refers to the process of first retrieving the most similar evidence from the legal corpus based on the user's question, and then providing this evidence as a reference for the existing LLMs to generate output. Retrieval-augmentation does not require any additional training for the existing LLM as it merely connects an

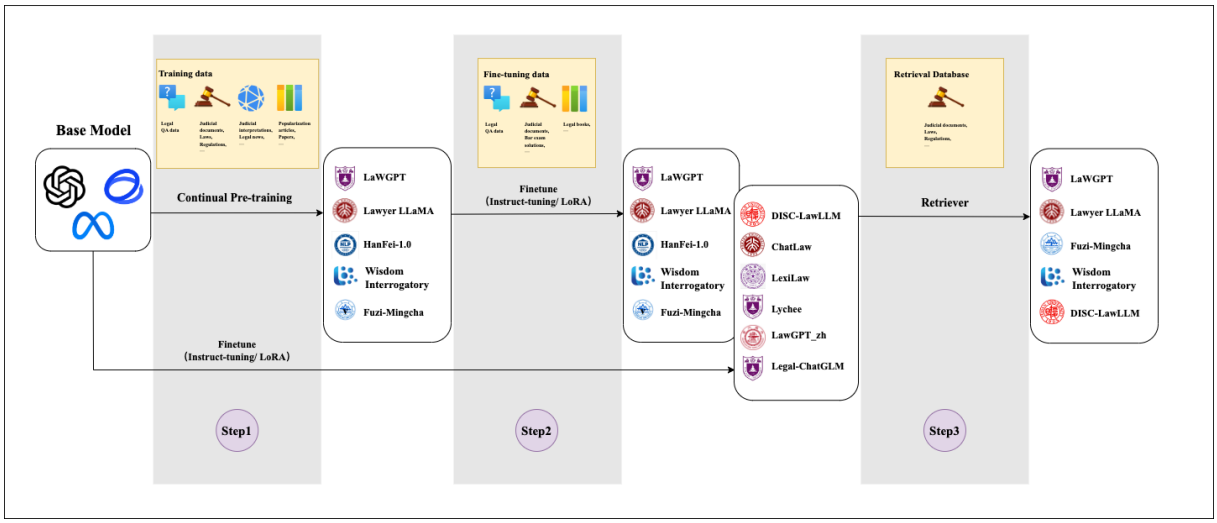


Figure 6: Legal LLM Training Process.

external knowledge base to the LLM. It's akin to giving the LLM an open-book exam, where the most relevant corpus from the retrieval library is inputted into the LLM, allowing it to reference the retrieved content for its response. The retrieval-augmented approach effectively alleviates the hallucination problem of LLMs and addresses the issue of knowledge updates.

H Evaluation Metrics Overview

Evaluation of legal LLMs is generally structured around the feature of the downstream tasks. As shown in Table 9, tasks can be divided into two categories:

(1) Generation Tasks: These tasks require models to generate text (e.g., summarization or legal reasoning). The quality of generated outputs is typically assessed using metrics such as ROUGE-L (Steffes et al., 2023; Mullick et al., 2022) and BERT-Score (Kumar et al., 2024; Benedetto et al., 2023; Joshi et al., 2024), which measure the semantic similarity or correlation between the model's output and the reference answer. With the advancement of LLMs' comprehension capabilities, some evaluation tasks have begun incorporating LLMs as judges to assess performance from multiple dimensions (Cui et al., 2023a; Li et al., 2024c). For example, the competition of CAIL-2024 (Challenge of AI in Law) introduced a subjective evaluation metric in the legal consultation track[114], which involves simulated scoring by large models. This includes the coherence of generated dialogues, which evaluates if the answers are relevant, and the accuracy of legal knowledge, which checks if the legal references and provisions are correct.

(2) Decision Tasks: These tasks involve classification or extraction, such as multiple-choice legal question answering (Pahilajani et al., 2024), legal case retrieval (Feng et al., 2024; Padiu et al., 2024), or judgment prediction (Wu et al., 2023; Wang et al., 2024). In this setting, evaluation metrics often include Recall, Accuracy, F1 scores, Mean Absolute Error (MAE), as well as ranking-based metrics such as NDCG@k (Wang et al., 2013) and Mean Reciprocal Rank (MRR) (Wu et al., 2011). The detailed descriptions and calculation methods for the metrics are shown in Table 8. For instance, in the 2023 CAIL competition's case retrieval track[115], models are required to retrieve the 30 most relevant cases from a candidate pool of over 55,000 cases with performance measured by NDCG@30.

In addition, some evaluation frameworks incorporate supplementary tasks assessing model safety and performance. For example,*Evaluation Metrics and Assessment Methods for Large Legal Models (Draft for Comments)* (System et al., 2023) evaluates the first response time, concurrency, processing efficiency. SUPERLAWBENCH[116] proposed assessment tasks on national security, public security, ethics and morality. Although these factors are crucial for product-level deployment, they are generally less directly related to the core legal reasoning tasks.

[114] http://cail.cipsc.org.cn/task_summit.html?raceID=4&cail_tag=2024

[115] http://cail.cipsc.org.cn/CAIL-dataset.html

[116] https://data.court.gov.cn/pages/modelEvaluation.html

Table 9: Overview of Evaluation Metrics for LLMs in Legal Tasks

	Generation Tasks	**Decision Tasks**
Example Tasks	Summarization, fact extraction, legal reasoning generation, dialogue generation, etc.	Multiple-choice questions in legal exams, case retrieval, opinion alignment, etc.
Evaluation	Semantic similarity, relevance, LLM as Judge, etc.	Accuracy, recall, precision, ranking effectiveness, etc.
Common Metrics	ROUGE, BERT-Score, LLM-Score, etc.	F1-score, micro-F1, exact match (EM), NDCG@k, MRR, etc.

I Legal LLM Applications

Legal LLMs are designed to assist human in accomplishing legal tasks. As shown in section 2.2, different LLMs perform differently on legal tasks. Therefore, there is no single best legal LLMs; each LLMs has its strengths and weaknesses when applied to legal tasks. Consequently, legal professionals need to play a crucial role in the application of models in legal scenarios. In this section, we identify the beneficiaries of legal LLMs and discuss the potential usages and issues of LLMs in practice.

I.1 Applications to Judges

LLMs can enhance judicial efficiency, ease judges' workload, and foster fairness and justice. However, their courtroom implementation faces hurdles including public trust, promoting good governance, and avoiding biases. The public trust issue is paramount, as illustrated by a controversial case in Colombia where a judge used AI tool ChatGPT(Taylor, 2023), despite its successful execution of secretarial tasks. A study suggests decision-makers might exhibit "selective adherence" to AI suggestions aligning with their stereotypes(Alon-Barkat and Busuioc, 2023), potentially disadvantaging certain citizens. However, with transparency, LLMs can assist judges in non-decision-making tasks without necessarily damaging public trust.

To address these challenges, we need innovative, reliable, and secure legal LLM application patterns. Some governments and organizations have issued policies and recommendations which could form the basis for addressing these issues(onAI Good Governance , OxCAIGG). For instance, China's Supreme People's Court issued "Opinions on Standardizing and Strengthening the Judicial Application of AI" in 2022[117], positioning AI as an auxiliary tool in judicial work and upholding users' right to self-determination. Similarly, the UK Courts and Tribunals Judiciary released "AI Guidance for Judicial Office Holders" in 2023, outlining key risks and suggestions for AI use in courts, and clearly listing tasks recommended and not recommended for AI. These documents provide guidance, but legal-level implementation remains a future task.

I.2 Applications to Lawyers

LLMs have reshaped the way of understanding and acquiring knowledge, bringing significant benefits to the work of lawyers in various professional levels and business area. A litigation lawyer from a Sydney law firm, in the process of using ChatGPT, believes that ChatGPT has already reached the working capability of a first-year lawyer(Purtill, 2023). If a lawyer has a good understanding of business details and knows how to engineer the appropriate prompts, they can more fully leverage the advantages of LLMs to obtain more valuable outputs.

In addition to improving the efficiency of lawyers, factors such as client demands and trust may lead to the inevitable choice of using LLMs for lawyers. As the application capability of LLMs is expected to become one of the professional skills for lawyers in the future, and lawyers lacking this capability may be uncompetitive in the market. In other words, clients will not want stand-alone lawyers who eschew AI(ChatGPT and Perlman, 2022).

LLMs can be used for many tasks such as contract review, due diligence, document draft, case summary, cross-examine questions, evidence strengthen suggestions, etc.(Perlman, 2023)(Tan et al., 2023)(Noonan, 2023), enabling lawyers to concentrate on core tasks and deliver more cost-effective solutions to their clients. Currently, a significant number of law firms and legal tech companies have publicly announced their integration of the LLM for specific use cases(Shaver, 2024). For example, Allen & Overy (A&O), the leading international law firm, has integrated Harvey, built on a version of OpenAI's latest models enhanced for legal work, into its global practice(Iu and Wong, 2023). Harvey has been used for legal tasks such as contract analysis, due diligence, litigation, and regulatory compliance, improving the efficiency of A&O's lawyers[118]. According to the characteristics of lawyers' work, we categorize the current applications for lawyers into three parts: common functions, functions for litigation lawyers, and functions for non-litigation lawyers (Figure 7). It should be noted that the categorization here is primarily for the convenience of discussion, as the identity of a lawyer may not be unique in practice.

First, basic applications refer to the functions commonly used by all lawyers in their professional activities. However, using LLMs solely for legal article retrieval is impractical, as the content generated by LLMs based on probability is generally not

[117] https://www.chinacourt.org/article/detail/2022/12/id/7057666.shtml

[118] https://www.harvey.ai/

Table 10: LLM applications for judges

Functions	Details	Advantages
Verdict warning	LLMs can analyze the judgment standards of similar cases. When a judge'sverdict deviates from the standards, an alert will be issued, notifying the iudge to re-examine the case.	Promote fairness and justice.
Case distribution	In the stage of case-filing, LLMs will identify the complexity of a case, Then for simple case, LLMs can provide intelligent legal consulting services, direct the parties to the mediation platform or advise them to apply for the summary procedure.	Alleviate the pressure on judges' workand promote diversified resolution ofdisputes.
Legal document draft	Drafting legal documents automatically based on templates, evidence materials, relevant cases, and the judges' verdicts.	Improve work efficiency

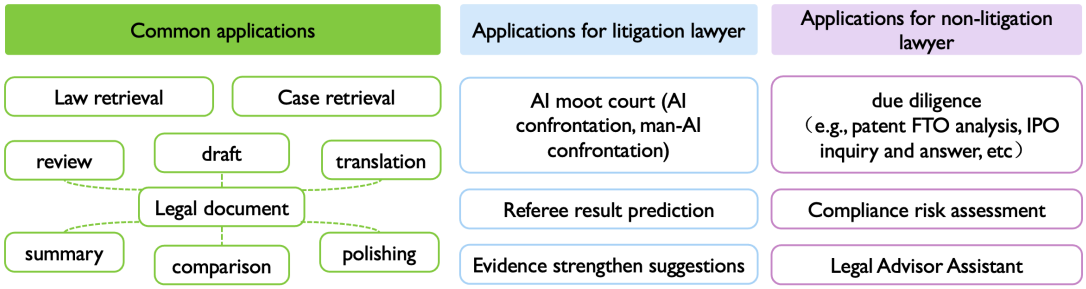


Figure 7: Examples of LLM applications in lawyers' work

real cases and laws. LLMs may need to integrate with information retrieval technology to fulfill these functions. In addition to these applications, LLMs have demonstrated good performance in tasks such as drafting, translation, review, summarization, polishing, etc. However, there may be certain issues in LLMs outputs. For instance, when using ChatGPT translate "overlapping of laws" into Chinese, the result is "duplicate laws", which is only a surface-level translation and doesn't match the existing legal terms in Chinese. Another example is that in analyzing bona fide acquisition system, ChatGPT believes someone is unable to obtain the ownership if he/she fails to return the subject matter maliciously which have been bought in good faith. It is evidently inconsistent with legal provisions.

Second, applications for litigation lawyers include AI moot court, referee result prediction, evidence strengthen suggestions, etc. Litigation lawyers typically engage in activities such as litigation and arbitration, involving tasks such as client meetings, evidence collection, and court debates. The mentioned applications can provide convenience for their work. Take AI moot court as an example, a lawyer can instruct LLM "The following are the relevant facts of the case: xxx. You are the defense attorney, and the user you are conversing with is the plaintiff's attorney. Please simulate a court debate, discussing the facts, procedures, etc., of the case." AI moot court is a typical application scenario based on LLM multi-turn dialogues. During the conversation, AI can effectively assist the lawyer in quickly identifying potential shortcomings in the regulations, evidence, argument points, and debate strategies they have prepared, thereby increasing the likelihood of a successful outcome. Some legal tech companies, such as PKULAW, have implemented AI moot court[119]. In addition to court debates, it also provides features such as simulated judgments and intelligent legal provision references. However, in real litigation, there may be many unpredictable situations, such as the evidence provided by the opposing party in court, the judge's attitude, and changes in defense strategy. The handling of all these situations requires strong observational skills, adaptability, and reliance on the experience and wisdom of lawyers.

Third, LLMs can provide support for non-litigation lawyers in due diligence, compliance risk assessment, legal advisor assistant, and other tasks. For instance, IP lawyers conducting Freedom to Operate (FTO) analysis need to comprehensively evaluate existing patents in a specific technical field

[119] https://ai.pkulaw.com/gpt/courtchat

to help businesses avoid unintentional infringement of others' patent rights. Lawyers often spend a significant amount of time on FTO analysis. On one hand, patent documents contain numerous technical terms, requiring lawyers to understand the technology before conducting the analysis. On the other hand, the global patent filing volume reached 3,457,400 in just 2022(, WIPO), not to mention the accumulated number of patents. Therefore, even for patents in a specific subfield, the quantity of documents requiring comparative analysis is quite substantial. LLMs can facilitate FTO analysis in several ways. Firstly, LLMs can rapidly interpret the technical content in patent documents and provide concise, easily understandable output to the lawyer. Secondly, LLMs excel at performing text comparative analysis, and lawyers can input the content they need to compare into LLMs to obtain analysis results. Thirdly, in the future, LLMs could further enhance its assistance in FTO analysis by enabling batch import of patent files, using patent files as training data for post-training or fine-tuning, subdividing patent technologies, and automating the generation of analysis reports.

Overall, there are many LLMs applications for lawyers and some of them have already performed well on common tasks such as contract review and due diligence. However, some applications have been developed but the effectiveness still needs to be improved, such as in applications for litigation lawyers, LLMs still struggle to take into account many factors in reality. For lawyers, as LLMs become more sophisticated and capable of performing complex legal tasks, they need to consider assigning simple, repetitive tasks to LLMs and focus more on areas where they can add the most value. Constructing clear and suitable prompts is a prerequisite for lawyers to effectively utilize LLMs. Legal Prompt Engineering (LPE) refers to the construction techniques of prompts in legal field. Research by Dietrich Trautmann(Trautmann et al., 2022) has found that even for LLMs that have not post-training or fine-tuning with legal data, they can still perform well in legal Q&A scenarios by constructing appropriate prompts. This highlights the importance of LPE. Currently, some companies or legal practitioners have actively made attempts in this aspect and have released a series of LPE cases for reference. For example, legal tech company CaseMark has published an LPE guide, providing examples in various aspects to illustrate the principles that lawyers should follow when constructing prompts[120].

Lawyers, as providers of legal services, should be vigilant about potential issues such hallucinations, copyright infringement, privacy breaches, discrimination, etc., when using LLMs. LLMs are not infallible and require professional and responsible supervision by lawyers(Murray, 2023). Otherwise, they may be liable for legal consequences arising from these issues. For example, Steven Schwartz, a lawyer from the New York law firm Levidow, Levidow & Oberman, with over 30 years of legal practice in the United States, incurred a $5,000 fine after being discovered by a judge for directly citing six cases collected by ChatGPT in a litigation case without verifying the authenticity(Howlett and Sharp, 2023). To avoid risks of using generative AI in legal services, the State Bar of California has published Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law, which is centered on the existing professional responsibility obligations for lawyers and illustrates how to act in accordance with these obligations. For example, to fulfill the duties of competence and diligence, lawyers need to not only review the outputs of LLMs for issues like false, inaccuracies or bias, but also gain an in-depth understanding of LLMs' working principles, limitations, terms of use, and policies regarding client data in advance. In the absence of common industry standards, the guidance can provide clear recommendations for lawyers to use AI in compliance with regulations.

I.3 Applications to Law School Students

LLM as a research subject

LLMs significantly reduce the barriers to utilizing and studying Artificial Intelligence (AI), enabling law students without technical backgrounds to engage with advanced technologies.

Before LLMs, traditional language models like LEGAL-BERT(Chalkidis et al., 2020) and Lawformer(Xiao et al., 2021) were employed to process large-scale legal data. AI automates some routine legal tasks and replaces some lower-rung legal functions such as basic memo checking and case drafting(Vučić, 2023; Alarie et al., 2018), and discussions on the legal governance of AI technology are frequent. All these developments necessitate

[120] ttps://www.casemark.ai/post/introduction-to-legal-prompt-engineering

a clear understanding of AI mechanisms among law students. While many law schools discuss introducing courses on AI and its legal implications (Goldsworthy, 2020), the study LAW2020[121] shows that only one in five has already incorporated these classes into their curriculum. This is largely due to the complexity of mastering traditional AI mechanics for law students. Without a technical background, they face a steep learning curve to master dataset cleaning, model training, and downstream tasks like contract analysis and case retrieval. Even with a trained model, developing an AI system that performs legal tasks can be a daunting task for law students.

However, LLMs, with their uniform input/output format and user-friendly prompt-tuning mechanism, are more accessible. An average user can easily understand how to use LLMs like ChatGPT and comprehend the prompt construction and associated risks. Through some basic learning of LLMs, law students can dialogue with LLMs and analyze the responses from a legal perspective, such as whether the output answers adhere to the legal reasoning syllogism, and whether there are fabricated laws, fictional facts, and other issues. Furthermore, they can evaluate the robustness of LLMs under different legal scenarios with different prompts, and even build specific legal LLMs. After mastering the fundamentals of LLMs, law students can provide professional support for policy-making, AI ethics evaluation, and intelligent legal governance.

LLMs have certain limitations as research subjects. LLMs are black boxes, easy to use but hard to understand in-depth. Due to the large number of parameters, the cost of training with full parameters and deploying is very high. These challenges make it difficult for law students to engage with and understand AI at a higher level. While they can operate the model, evaluate the output, and comprehend issues like legal hallucinations, how to optimize remains a hurdle. This partial knowledge may lead to biased judgments when analyzing and resolving legal issues related to LLMs.

LLM as a learning tool

LLMs can serve as a supplementary tool in the daily study of law students. Before the emergence of LLMs, other AI technologies have been found to have some potential in facilitating legal education, including legal retrieval, information extraction, and outline drafting, etc. For example, some law schools employ automatic extraction models to aid students in Case-Based Argumentation(Aleven and Ashley, 1997). However, Law2020 shows most law professors concurred that AI has not yet fully emerged as a meaningful teaching tool. This could be attributed to earlier AI applications being deficient in accuracy and reusability and associated with high maintenance costs. LLMs largely address these issues.

LLMs nowadays are more accurate than previous models. When taking the bar exam, GPT-3.5 scored in the bottom 10th percentile, while GPT-4 not only passed but also scored in the top 10th percentile(Katz et al., 2024). Furthermore, LLMs are highly reusable. In the past, specialized models had to be individually designed for each task. However, a legal LLM can accommodate a broad spectrum of learning requirements, like generating summaries and extracting legal elements(Kasneci et al., 2023). Moreover, LLMs are easier to update and maintain. In practice, laws and judicial applications change frequently, necessitating regular updates to textbooks and other reference materials. Traditional AI models require re-training if data changes, but LLMs can ensure up-to-date information by updating the external knowledge base for retrieval-augmentation or implementing forgetting strategies.

More importantly, compared with previous AI teaching tools, LLMs can provide information as well as explanations and analyses. For example, when revising legal documents written by students, LLMs can do more than correct grammatical and formatting errors. They can also provide suggestions for refining and enhancing the document. When providing case briefs, LLMs can not only extract legal facts and conclusions, but also help analyze the judgment logic and the sentencing factors(de Faria et al., 2024; Yue et al., 2023).

Beyond being a practical teaching tool, LLMs introduce new learning paradigms. For example, in the Socratic Playground for Learning System(SPL)(Zhang et al., 2024b), LLMs foster critical thinking by posing questions and gradually breaking down complex problems, enabling tailored learning scenarios and efficient multi-turn tutoring dialogues. This approach is particularly well-suited to law school courses such as Moot Court and Legal Negotiation, where logical reasoning and analytical thinking are paramount. By

[121] https://abovethelaw.com/law2020/cognifying-legal-education/

enabling role-playing exercises that simulate real-world legal scenarios, LLMs can enhance students' practical skills and professional competencies.

Currently, attempts have been made to introduce LLMs into the law school classroom. Lexis+AI(Mika, 2022), a tool that supports conversational search, intelligent legal drafting, insightful summarization, and document analysis, is set to be accessible to 100,000 law students in the 2024 spring semester[122]. However, LLMs are not perfect as teaching assistants. The output of LLMs is not entirely reliable, especially in specific laws or legal concepts. Although Lexis+AI mitigates legal hallucination by linking legal citations, the expression may change during the generation process, resulting in a lack of precision. This can be particularly challenging for beginners to discern the quality of the output. As a result, Lexis+AI is only available to senior law students and faculty. In addition, an over-reliance on LLMs could potentially hinder students' growth and suppress their creativity.

Therefore, LLM tools should be used with legal professional oversight. Students should be informed about the risks of using LLMs and encouraged to study security, privacy, bias, and other issues.

I.4 Applications to the General Public

Accessing legal information is challenging for the general public. Financial constraints, fear of the law, and complex procedures often hinder the average person from seeking legal advice. Despite government's efforts to expand public legal aid, these measures often fail to meet individual needs(Mansfield and Trubek, 2011).

Before the advent of LLMs, digital tools were used to bridge the gap between ordinary people and lawyers. These tools are based on rules or incorporate traditional AI techniques. To some extent, they reduce the cost of money when people seek legal help and expand the coverage of legal aid. And they allow users to ask questions without psychological burden, especially for private matters concerning marriage and family. However, these tools have limitations regarding their capabilities, usage thresholds, and practical application coverage.

Rule-based legal tools. These tools, designed by legal experts, aim to provide precise calculations for specific legal requirements but face challenges in reusability and accessibility for the average user. For instance, when calculating industrial injury compensation[123], legal experts need to gather information from various laws, regulations, departmental rules, and local provisions. They then integrate this information with relevant case studies to develop a complete calculation system, accounting for differences in regional and temporal calculation rules. If new injury regulations alter the logic, the entire process must be repeated. Similarly, developing tools for other scenarios, such as private lending claims, requires consulting experts to summarize and codify new rules from scratch, as these tools cannot be directly transferred. Additionally, these systems rely on logical decision-making, requiring users to answer questions about their specific legal situations. However, the use of technical legal terms, such as "the stage of industrial injury compensation", often creates barriers for non-experts, making it challenging for ordinary users to understand and effectively utilize these tools.

Traditional AI-based legal tools. These tools are typically AI combined with rules(Dias et al., 2022). They have two main applications: questionnaire-based legal advice generation and dialogue-based legal advice generation. Due to the limitation of model scale, they can only serve for certain services, and perform poorly on these limited applications, both of which have limitations in service diversity and accuracy. The tool performs as a case questionnaire first lets users select a type of consultation, such as counsel fee or brief fee[124]. After that, users fill out a corresponding case questionnaire about basic information and facts. Based on the questionnaire, the tool generates legal advice. However, the questionnaire-based method restricts users from asking questions or giving feedback, potentially failing to address their unique needs. Dialogue-based, like the AI assistant in the China Legal Service Website[125], recommends and formulates questions based on user input. However, they are incapable of engaging in comprehensive conversations. Once a user's question is not in the database, it can only respond with 'I don't know'. Although there are some improved methods like knowledge graphs(Sovrano et al., 2020), the

[122] https://www.abajournal.com/web/article/law-students-gain-access-to-lexis-ai-generative-ai

[123] https://www.hshfy.sh.cn/shfy/fzgj/c2jtools/yibgspc.html

[124] https://www.fagougou.com/pc/?mkt=fidml0542ab94

[125] http://www.12348.gov.cn/#/homepage

graphs' limited information can only cover certain scenarios.

LLMs inherit the advantages of previous legal tools in convenience, while also addressing their shortcomings in accuracy and adaptability. Compared with previous tools, LLMs have shown substantial advancements in intelligence. On the one hand, LLMs enable users to interact with the model without any prior selection of consultation areas. This means that LLMs are not restricted to some high-frequency legal scenarios, and users can set up personalized dialogue scenarios to accommodate a wide range of legal needs. On the other hand, LLMs bridge the comprehension gap in legal terminology that the general public faces. Unlike previous tools, LLMs can interpret legal terms in a common language for questioners, avoiding professional jargon. If users still find certain words challenging to understand, they can query the LLM repeatedly until they grasp the meaning. The extensive pre-training data enables LLMs to offer reasonably accurate responses based on their experience, even in unfamiliar situations.

LLMs encompass the application scenarios covered by all the previous legal tools, offering more comprehensive support. With a rich pre-trained corpus and external knowledge base or tools, LLMs can supply abundant supporting material. For example, when users consult government information, LLMs can retrieve relevant information from government websites. When explaining laws and regulations, LLMs can cite relevant cases and judicial interpretations[126]. More significantly, unlike the passive reception of instructions from previous tools, the LLM functions more akin to a server, delivering thoughtful service based on experience, even when users are unfamiliar with their legal situation. The average person, for example, may only have a vague awareness of potential legal risks, without understanding how to address them. LLMs can restructure the case based on the user's description, clearly inform the user of the current legal situation, provide subsequent coping strategies (such as rights protection or prosecution), and match a template to assist the user in writing legal documents.

Current LLMs are still in development, with many shortcomings limiting their large-scale application in legal aid. Regarding accuracy, LLMs may struggle to understand and respond appropriately to unclear user queries. Moreover, these models struggle to effectively utilize legal knowledge despite undergoing legal pre-training and having access to external knowledge bases. For example, they often fail to correctly apply laws and regulations in specific regions, and are hard to promptly disregard obsolete laws. As generative models, LLMs also lack the ability to perform precise numerical calculations, and their propensity for generating inaccurate information could potentially mislead laypeople. To improve the accuracy of advice given, LLMs could incorporate rule-based patterns. For instance, using the user's background information (such as gender, age, region, etc.) as a part of prompts to guide generation, or invoking additional calculation tools when dealing with numerical calculations such as monetary amounts or prison sentences.

While LLMs can offer guidance during the initial stages of legal aid, the assistance of professional lawyers is indispensable as the legal process progresses. In the early stage, users' primary needs are to understand their situation and clarify the next steps, such as informing litigation procedures, drafting the complaint, and guiding the evidence collection. Even if the information provided by LLMs is not entirely precise, it is generally acceptable. However, in the later stages like trial and appeal, the application methods for specific cases vary, necessitating more professional and personalized services that current LLMs are unable to provide. In practice, users should be advised not to rely solely on these large models, and to seek further assistance from a lawyer if they have any doubts.

[126] https://tongyi.aliyun.com/farui