information

MDPI

Systematic Review

To What Extent Have LLMs Reshaped the Legal Domain So Far? A Scoping Literature Review

Bogdan Padiu [1], Radu Iacob [1], Traian Rebedea [1,2] and Mihai Dascalu [1,3,*]

[1] Computer Science & Engineering Department, National University of Science and Technology POLITEHNICA Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania; bogdan.padiu@stud.acs.upb.ro (B.P.); radu.iacob@upb.ro (R.I.); traian.rebedea@upb.ro (T.R.)

[2] NVIDIA, Santa Clara, CA 95051, USA

[3] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania

[*] Correspondence: mihai.dascalu@upb.ro

Abstract: Understanding and explaining legal systems is very challenging due to their complex structure, specialized terminology, and multiple interpretations. Legal AI models are currently undergoing drastic advancements due to the development of Large Language Models (LLMs) that have achieved state-of-the-art performance on a wide range of tasks and are currently undergoing very rapid iterations. As an emerging field, the application of LLMs in the legal field is still in its early stages, with multiple challenges that need to be addressed. Our objective is to provide a comprehensive survey of legal LLMs, not only reviewing the models themselves but also analyzing their applications within the legal systems in different geographies. The paper begins by providing a high-level overview of AI technologies in the legal field and showcasing recent research advancements in LLMs, followed by practical implementations of legal LLMs. Two databases (i.e., SCOPUS and Web of Science) were considered alongside additional related studies that met our selection criteria. We used the PRISMA for Scoping Reviews (PRISMA-ScR) guidelines as the methodology to extract relevant studies and report our findings. The paper discusses and analyses the limitations and challenges faced by legal LLMs, including issues related to data, algorithms, and judicial practices. Moreover, we examine the extent to which such systems can be effectively deployed. The paper summarizes recommendations and future directions to address challenges, aiming to help stakeholders overcome limitations and integrate legal LLMs into the judicial system.

Keywords: scoping review; legal datasets; large language models (LLMs); legal; judicial data; natural language processing

check for updates

Citation: Padiu, B.; Iacob, R.; Rebedea, T.; Dascalu, M. To What Extent Have LLMs Reshaped the Legal Domain So Far? A Scoping Literature Review. *Information* **2024**, *15*, 662. https://doi.org/10.3390/info15110662

Academic Editor: Katsuhide Fujita

Received: 10 September 2024
Revised: 4 October 2024
Accepted: 16 October 2024
Published: 22 October 2024

1. Introduction

Since November 2022, with the launch of ChatGPT, and especially after GPT4 [1], there has been an exponential increase in the use of LLMs in various research fields, including code generation, economics, healthcare, and education. Amidst this surge of innovation, the legal domain has remained particularly challenging. Legal systems are hard to understand and explain because of their complicated structure, specialized language, and varying interpretations. This complexity makes it difficult for both the general public and professionals to navigate the legal field [2]. Although technological advancements have the potential to create LLMs that simplify these systems, their effectiveness has not yet been proven, and multiple challenges remain. The multi-level hierarchies, domain-specific vocabulary, and nuanced interpretations inherent to legal matters pose significant challenges for these models. Consequently, the outputs generated by the models do not consistently provide the necessary depth and precision to provide meaningful assistance in real-world legal scenarios [3].

Nevertheless, the emergence of LLMs has introduced new opportunities for legal research and practice. These technologies can potentially enhance the accessibility, efficiency, and accuracy of legal information retrieval and processing. However, while the general application of LLMs can provide benefits, it also highlights the need for specialized systems tailored to the legal domain to realize their full potential. As is the case in code generation tasks [4], the unique nature of legal tasks makes common prompting techniques optimized for natural language tasks [5,6] significantly less effective. Off-the-shelf LLMs struggle to fully capture the complexities and nuances of legal language and reasoning [7,8]. They may fumble with the specialized terminology and citation formats or outright hallucinate domain-specific knowledge, losing the rigor and precision essential in legal contexts. To truly leverage the capabilities of LLMs in the legal field, it is necessary to develop models that are fine-tuned and adapted to the specific requirements of legal research and practice [9,10]. This involves training LLMs on extensive collections of legal texts, incorporating domain-specific knowledge, and optimizing them for tasks such as legal document retrieval, summarization, and analysis. Specialized legal LLMs can better understand the context and meaning of legal language, handle the unique structures and formats of legal documents, and provide more accurate and relevant results [9].

The traditional judiciary system possesses several distinct characteristics that are essential to comprehend when considering the application of judicial artificial intelligence [11]. These characteristics encompass a reliance on human decision-making, a lack of flexibility, and substantial resource consumption.

One of the primary features of the traditional judiciary is its dependence on human decision-making, particularly that of judges, prosecutors, and lawyers [8]. Throughout the process of reasoning and evidence collection, legal professionals often refer to case-specific circumstances, legal provisions, and precedents, in conjunction with their professional knowledge, to formulate judgments and decisions. The final judgment or defense is then presented through a trial.

Another key aspect of the traditional judiciary is its reliance on precedents during the decision-making process. Previous judgments in similar cases and relevant legal provisions often guide the decisions of courts. In many judicial systems, the judgments of the highest court are considered authoritative and binding, serving as a reference for other courts in relevant cases [8].

Furthermore, the traditional judiciary is often time- and resource-consuming, particularly when dealing with a large number of cases. The process of case hearings, summoning witnesses, and collecting evidence can prolong the trial process and consume significant resources [8]. This can lead to situations where there are many cases but limited personnel, resulting in backlog and delays.

Overall, while the traditional judiciary system relies on human expertise, legal norms, and precedents, it may lack flexibility and efficiency in certain situations. These characteristics highlight the potential for judicial artificial intelligence to streamline processes, enhance consistency, and improve access to justice while emphasizing the need for careful consideration of ethical and social implications.

As of the time of this review, there has been a growing number of start-ups that aim to offer solutions based on the industrial use of LLMs in the legal field. The available information regarding those solutions is rather scarce at the moment being mainly limited to marketing materials. We can infer, however, that there is growing confidence in the capability of LLMs to solve useful tasks in the legal domain. Some of the more well-known legal AI startups are Harvey (https://www.harvey.ai/, accessed on 29 September 2024), Lawpath (https://lawpath.com.au/, accessed on 29 September 2024), CaseText (https://casetext.com/, accessed on 29 September 2024), Robin AI (https://www.robinai.com/, accessed on 29 September 2024), Hopkins (https://www.hopkins.systems/, accessed on 29 September 2024), Henchman (https://henchman.io/, accessed on 29 September 2024), Juro (https://juro.com/, accessed on 29 September 2024), LegalFly (https://www.legalfly.ai/, accessed on 29 September 2024), Maigon (https://maigon.io/, accessed on 29 September 2024), Orbital Witness (https://www.orbitalwitness.com/, accessed on 29 September 2024), and RightHub (https://righthub.com/, accessed on 29 September 2024).

Research Questions

This scoping review seeks to explore five primary inquiries regarding the application of LLMs within the legal field, specifically focusing on their operational use cases and the methodologies employed in their construction:

- RQ1: Which LLM tools are considered leading in the field, and which are best suited for legal applications according to the current open-access state-of-the-art research?
- RQ2: What are the primary sources for data extraction and the best strategies for dataset development within the legal domain?
- RQ3: What are the challenges of LLMs in addressing legal tasks?
- RQ4: What are the main strategies for increasing the performance of LLMs in addressing legal tasks?
- RQ5: What are the main limitations of current LLMs for the legal domain?

2. Method

Our study considered the PRISMA extension for Scoping Reviews (PRISMA-ScR) [12] and PRISMA 2020 [13] methodology to provide syntheses of the state of knowledge in legal applications based on artificial intelligence, from which future research priorities can be identified. The study protocol was registered on INPLASY (INPLASY202490055).

2.1. Eligibility Criteria

We selected studies concerning the application of LLMs in the legal domain. Since the speed with which the domain has evolved in the past years is very high, we considered only recent works to establish the current state-of-the-art: newer than 1 January 2022 and up until 28 March 2024. We used mainly Scopus and Web of Science databases to calibrate the search criteria and decided early on that we would focus only on articles published in English.

On Web of Science, we refined the search to include only open-access studies. For the rest of the sources, we tried to retrieve the article, and when not successful, we excluded the article. For the selected and analyzed articles, we also included in the analysis any cited papers that satisfied the criteria. An overview of the selection process can be seen in Figure 1.

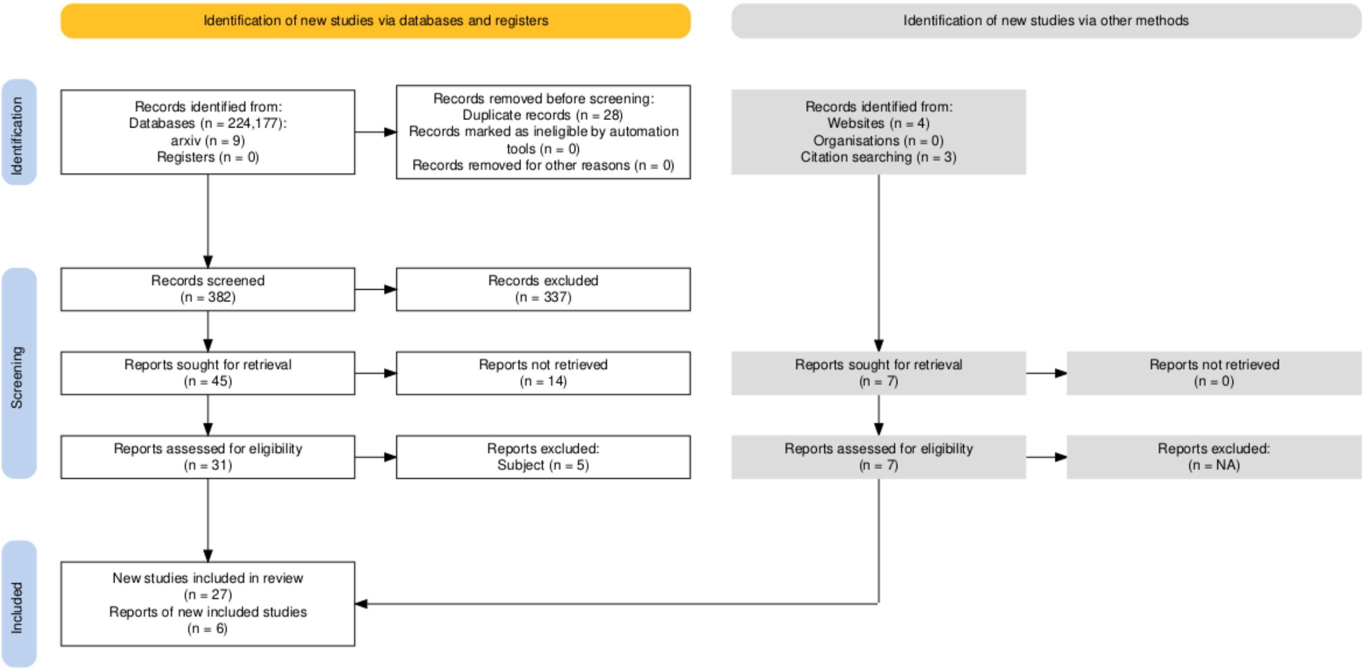

Identification of new studies via databases and registers
Identification of new studies via other methods
Identification
Screening
Included
Records identified from: Databases (n = 224,177): arxiv (n = 9) Registers (n = 0)
Records removed before screening: Duplicate records (n = 28) Records marked as ineligible by automation tools (n = 0) Records removed for other reasons (n = 0)
Records identified from: Websites (n = 4) Organisations (n = 0) Citation searching (n = 3)
Records screened (n = 382)
Records excluded (n = 337)
Reports sought for retrieval (n = 45)
Reports not retrieved (n = 14)
Reports sought for retrieval (n = 7)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 31)
Reports excluded: Subject (n = 5)
Reports assessed for eligibility (n = 7)
Reports excluded: (n = NA)
New studies included in review (n = 27) Reports of new included studies (n = 6)

Figure 1. The corresponding PRISMA diagram [14], depicting our selection process.

2.2. Search Strategy

The initial search results formed a starting set of papers that underwent analysis. During this analysis process, we identified additional related studies that aligned with the established selection criteria. These newly discovered relevant papers were then incorporated into the reviewed set, expanding the scope of the literature under consideration.

The search query used to retrieve a list of studies from the SCOPUS database on 22 February 2024 was as follows:

```
TITLE-ABS-KEY("LLMs" OR "*gpt*" OR "llm" OR "machine learning" )
AND TITLE-ABS-KEY("law" OR "legal" OR "contract")
AND PUBYEAR \(>\) 2021
AND SUBJAREA(SOCI )
AND (
LIMIT-TO ( SUBJAREA,"BUSI" )
OR LIMIT-TO ( SUBJAREA,"COMP" )
OR EXCLUDE ( SUBJAREA,"ENER" )
OR EXCLUDE ( SUBJAREA,"PSYC" )
OR EXCLUDE ( SUBJAREA,"HEAL" )
OR EXCLUDE ( SUBJAREA,"MEDI" )
OR EXCLUDE ( SUBJAREA,"CHEM" )
OR EXCLUDE ( SUBJAREA,"CENG" )
OR EXCLUDE ( SUBJAREA,"PHYS" )
OR EXCLUDE ( SUBJAREA,"BIOC" )
OR EXCLUDE ( SUBJAREA,"MATE" )
OR EXCLUDE ( SUBJAREA,"EART" )
OR EXCLUDE ( SUBJAREA,"ARTS" )
OR LIMIT-TO ( SUBJAREA,"MATH" )
)
AND (LIMIT-TO ( LANGUAGE,"English" ))
AND (EXCLUDE ( DOCTYPE,"ch" ) OR EXCLUDE ( DOCTYPE, "bk" ))
```

The analogous search query used to retrieve a list of studies from the Web of Science database on 22 February 2024 (See Table 1) was as follows:

```
(TI=("LLMs" OR "*gpt*" OR "llm" OR "machine learning"
OR "LQA")
OR AB=("LLMs" OR "gpt*" OR "llm" OR "machine learning" OR "LQA"))
AND (TI=("law" OR "legal" OR "contract") OR AB=("law" OR "legal" OR "contract"))
AND LA=(English)
NOT DT=(Book OR Book Chapter)
AND Arts Humanities Other Topics or Astronomy Astrophysics or Marine Freshwater
Biology or Materials Science or Mathematical Computational Biology or Obstetrics
Gynecology or Oceanography or Oncology or Ophthalmology or Optics or
Otorhinolaryngology or Pathology or Pediatrics or Urology Nephrology or Toxicology
or Physics or Pharmacology Pharmacy or Imaging Science Photographic Technology
or Immunology or Infectious Diseases or Instruments Instrumentation or Integrative
Complementary Medicine or Gastroenterology Hepatology or General Internal Medicine
or Genetics Heredity or Geochemistry Geophysics or Geography or Geology or
Geriatrics Gerontology or Electrochemistry or Cell Biology or Chemistry or
Cardiovascular System Cardiology or Biotechnology Applied Microbiology or
Biophysics or Biomedical Social Sciences or Biochemistry Molecular Biology or
Automation Control Systems or Anesthesiology or Acoustics or Physical Geography or
Physiology or Plant Sciences or Polymer Science or Psychology or Psychiatry or
Radiology Nuclear Medicine Medical Imaging or Rehabilitation or Remote Sensing or
Reproductive Biology or Research Experimental Medicine or Spectroscopy or Surgery
or Telecommunications or Thermodynamics or Virology or Veterinary Sciences
(Exclude - Research Areas)
AND Computer Science OR Mathematics (Research Areas)
AND Open Access
```

Table 1. Distribution of reviewed studies across different sources.

Source	Source Type	Date	Number of Studies
Web of Science	Database	22 February 2024	224
SCOPUS	Database	22 February 2024	177
Specific websites	Search	28 March 2024	4
arxiv.org	Register	28 March 2024	12
Total studies			417

2.3. The Selection Process

For this scoping review, we selected studies focusing on applying LLMs in legal contexts. Given the rapid evolution of this domain, we decided to only include very recent works published after 1 January 2022, to establish the current state of the art. We focused exclusively on English language articles to minimize risks associated with automatic translation.

Our primary databases for calibrating the search criteria and retrieving studies were Scopus and Web of Science. We developed complex queries, and iteratively, we tested several queries to achieve a balance between the number of returned results and the review capacity.

The final queries provided at the previous point resulted in 382 records. Each record was initially screened by one reviewer based on the title and abstract to assess its eligibility for inclusion in the review. The primary criteria were the target domain and the actual application of LLMs within the legal field. Articles that matched the query string and contained relevant keywords but focused on non-legal domains, such as healthcare, finance, or education, were excluded. For example, some articles addressed the legal aspects of AI in other industries but not the use of LLMs within the legal domain. In such cases, although the keywords matched, the semantic content did not align with the review's focus.

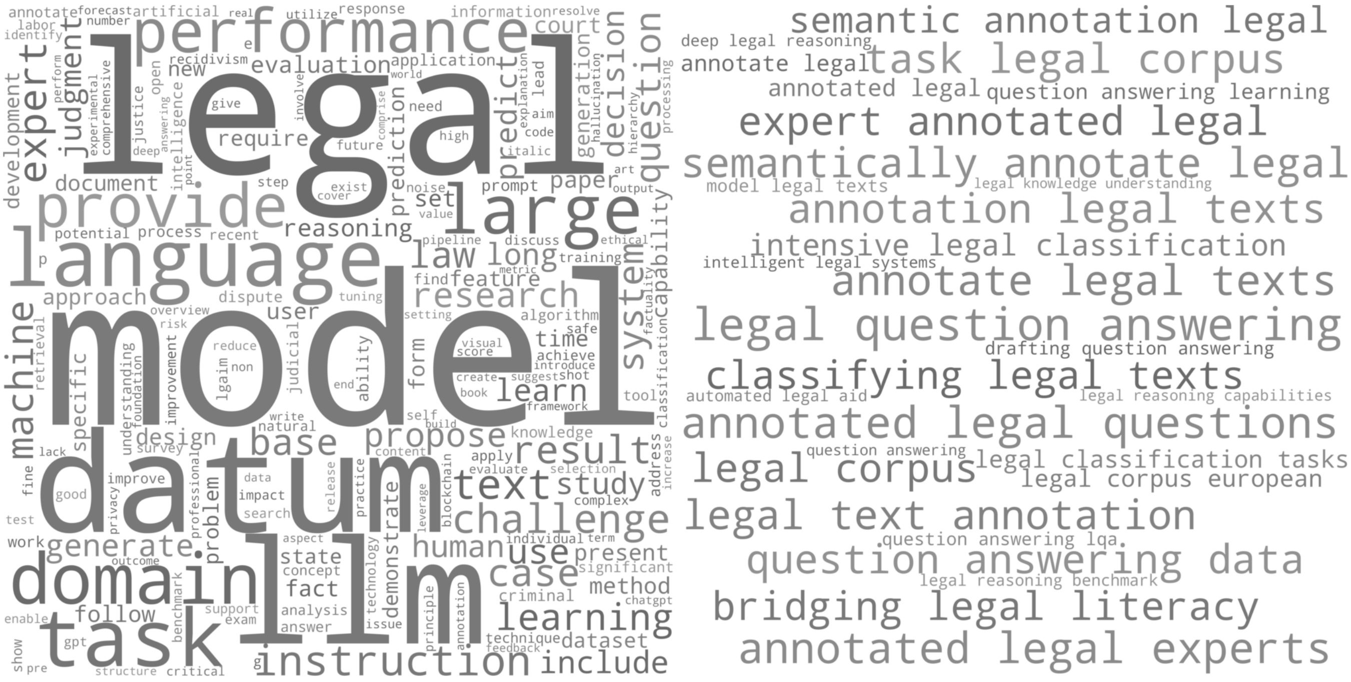
Out of the initial set of results, only 45 articles were eligible for in-depth analysis. In 14 instances, we were unable to retrieve the full text of articles due to access restrictions, as the papers were not available as open access. For each retrieved report, a high-level review followed to decide if indeed the research article satisfied the eligibility criteria. The first reviewer consulted a second, more experienced reviewer to validate decisions, thereby mitigating risk despite the absence of standardized coding protocols in this scoping review. Additionally, for the selected and analyzed articles, we also included any cited papers that satisfied our inclusion criteria in the analysis.

3. Results

3.1. Bibliographical Analysis

As part of our bibliographical analysis, we highlighted the key themes and focus areas within the selected body of literature. To achieve this, we conducted a keyword analysis using the final 33 selected research papers. We employed two automated keyword extraction tools: (a) frequency of content words (i.e., nouns, verbs, adverbs, and adjectives) using spaCy [15] and (b) KeyBERT [16]. First, we extracted the abstracts and the keywords from the .bib file containing the article references. Then, we processed the abstracts and the keywords with the appropriate tools. The first method generated a word cloud based on individual keywords extracted from the abstracts, offering a broad view of the general themes and terms prevalent across the papers. The second method (KeyBERT) focused on extracting the top 30 key three-word sequences, providing a more detailed perspective on specific topics and interrelationships within the research. Additionally, we created a word cloud using the keywords provided by the authors of each paper, providing insight into the authors' intended focus areas. This approach allowed us to examine the recurring concepts and emerging trends, contributing to a better understanding of the current research landscape.

The first word cloud (see Figure 2a) showcases the overall emphasis of the papers by highlighting broad terms like "legal", "model", "language", and "data", indicating the general focus on legal systems, LLMs, and their applications. Figure 2b is more balanced and narrows down to more specific phrases such as “legal question answering”, “annotated legal texts”, and “semantic annotation”, which highlights the detailed tasks and challenges being addressed in the research. This method reveals specific focus areas within the broader themes identified in the first word cloud.



(**a**) (**b**)

Figure 2. Automatically generated keywords from abstracts. (**a**) Keywords based on the frequency of content words. (**b**) Keywords extracted with KeyBERT.

The third word cloud based on author-defined keywords (see Figure 3) highlights terms like “machine learning”, “natural language processing”, and “artificial intelligence”, which are central to the technical aspects of the research. The presence of terms like “ethics”, “trustworthiness”, and “interpretability and explainability” indicates a broader concern with the implications and applications of these technologies.

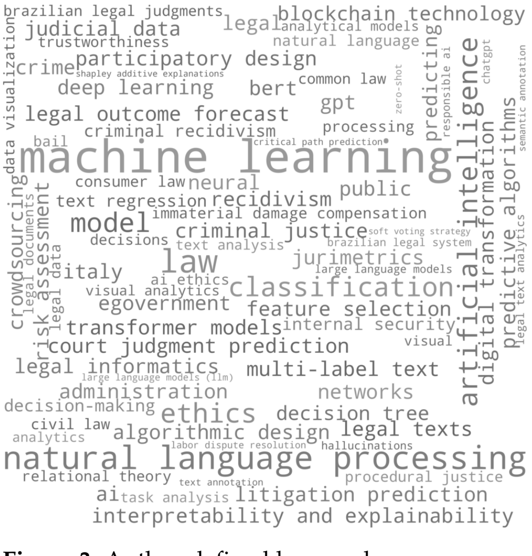


Figure 3. Author-defined keywords.

By comparing these word clouds, we observed how different keyword extraction methods (i.e., automated or author-defined) emphasized different aspects of the research.

Combining these approaches offered a high-level view, balancing the broad themes derived from abstracts with the specific areas of focus highlighted by both automated extraction and author intent.

3.2. Distribution Trends

Next, we analyzed the distribution and trends within the selected body of literature, and we conducted a statistical analysis that focused on three key aspects: geography, publication month, and publication channel. First, we analyzed the geographical distribution of the articles to identify the regions most active in research on the legal applications of LLMs. This geographic analysis helped pinpoint global centers of research and innovation within the domain. Additionally, we examined the number of articles published per month to uncover any temporal trends or spikes in research activity, providing insights into the evolution of interest and focus over time. Lastly, we analyzed the distribution of articles across various publication channels, including journals, conferences, and other platforms, to better understand the preferred avenues for disseminating advancements in this rapidly evolving field.

Figure 4 shows the distribution of included studies by geography. In order to assign a study to a geographical area, we analyzed the research organizations that authored the document. Analyzing the chart figures, we observed that the US + CA was the geography with the highest number of studies conducted, with Europe and China standing out as second and third in these research efforts.

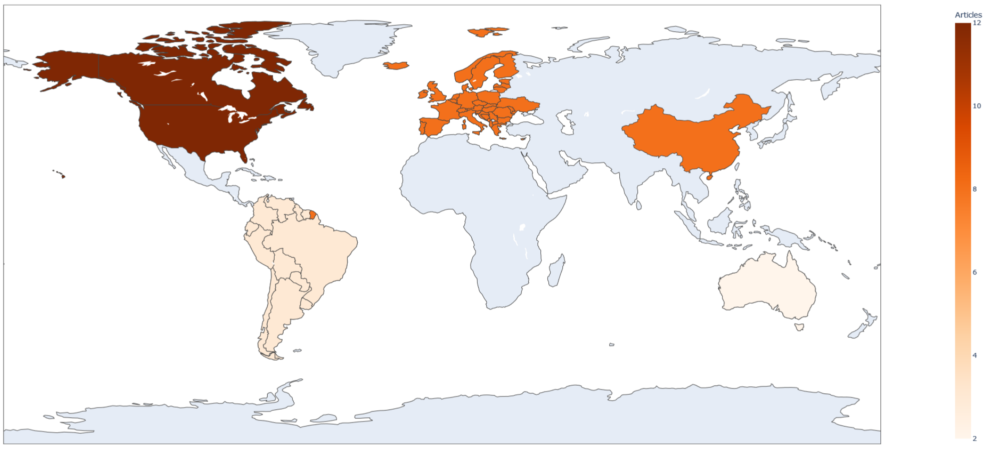


Figure 4. Publications per geography.

In Figure 5, we observe a clear increase in the number of publications concerning LLMs used in the legal domain from 2022 to 2024. The distribution of research on legal applications of LLMs across different publication channels reflected diverse dissemination strategies. A significant portion of studies were available on platforms like Arxiv, which are preprint repositories allowing researchers to share their work quickly before formal peer review. Other key channels included conference proceedings and journals such as Frontiers and ACM, demonstrating a strong preference for engaging both the academic community and wider audiences. This pattern highlights the balance between the need for rapid dissemination through preprint servers and the importance of publishing in established academic journals (See Figure 6).

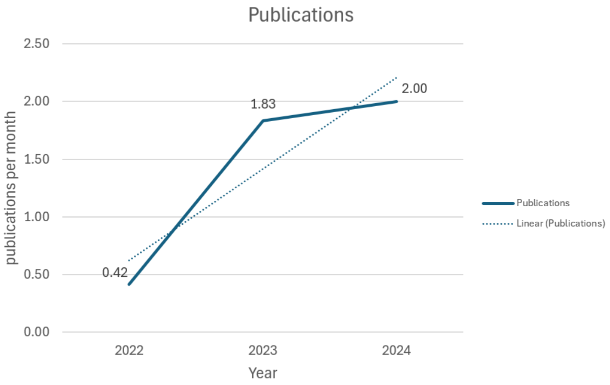


Figure 5. Publication frequency normalized per month. For 2024, we only considered the first three months.

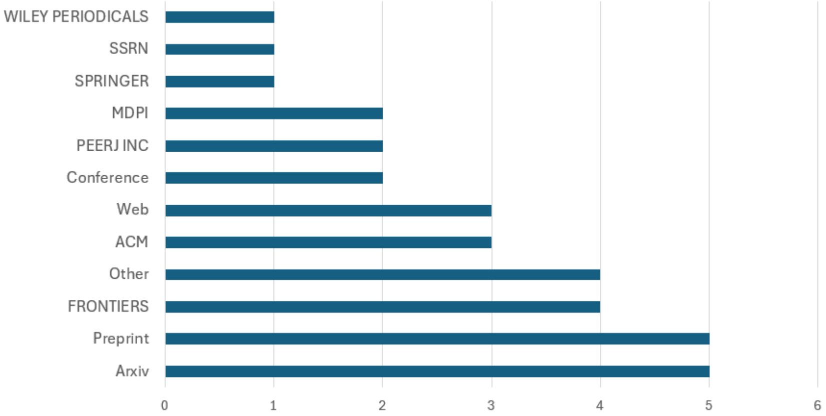


Figure 6. Publications per channel.

3.3. Main Events and Publications

The selected literature, the cited sources within the papers, and the reviewers' expertise enabled the identification of several key conferences and events central to advancing research on AI applications in the legal domain. The International Conference on Artificial Intelligence and Law (ICAIL), which held its 19th edition in 2023 (https://icail2023.di.uminho.pt, accessed on 1 May 2024), has been the leading global conference on AI and law since its inception in 1987. Organized biennially by the International Association for Artificial Intelligence and Law (IAAIL) and in cooperation with the Association for the Advancement of Artificial Intelligence (AAAI), ICAIL's proceedings are published by ACM, making it an important forum for the latest research in the field. As an associated event of ICAIL 2023, the 10th Competition on Legal Information Extraction and Entailment (COLIEE-2023) was also held, highlighting the latest advancements in legal information processing and providing a platform for testing and benchmarking state-of-the-art models in the field (https://sites.ualberta.ca/~rabelo/COLIEE2023/, accessed on 1 May 2024).

The International Conference on Legal Knowledge and Information Systems (JURIX), now at its 36th edition as of 2023 (https://jurix23.maastrichtlawtech.eu/, accessed on 1 May 2024), has been another key venue for exploring the intersection of law, AI, and Information Systems. Since 1988, JURIX, organized by the JURIX Foundation for Legal Knowledge Systems, has gathered researchers from the Netherlands, Flanders, and beyond, establishing itself as an important event in this interdisciplinary field.

The International Workshop on Juris-informatics (JURISIN), which held its 17th edition in 2023 (https://research.nii.ac.jp/~ksatoh/jurisin2023/, accessed on 1 May 2024), focuses on the study of legal issues through informatics. This workshop draws participants from law, social science, intelligent technology, and philosophy, creating a space for discussing both theoretical and practical challenges in AI and law.

The Annual Conference of the Society for Computers and Law (SCL) celebrated its 50th anniversary in 2023 (https://www.scl.org/13010-scl-50th-anniversary-conference-event-report-the-dragon-under-the-sofa-and-other-stories/, accessed on 1 May 2024). The event highlighted AI's transformative impact on the legal profession, with Professor Susskind describing the introduction of technologies like the GPT series as a "seismic moment". The conference also introduced the SCL AI Group's Artificial Intelligence Contractual Clauses, offering guidance and sample clauses for drafting and negotiating AI-related contracts (https://www.scl.org/wp-content/uploads/2024/02/AI-Clauses-Project-October-2023-final-1.pdf, accessed on 1 May 2024).

3.4. Main Legal Tasks

The selected literature revealed several key tasks for applying LLMs within the legal domain.

Legal case retrieval is a key task in case law systems, making this task particularly challenging and vital for developing intelligent legal systems. Given the complexity of legal texts and the importance of precedents, this task is both difficult and essential. The current SOTA is the THUIR team's work at COLIEE 2023 [17], where they won the competition with the Structure-Aware pretrained language model for LEgal case Retrieval (SAILER) (Task 1 Leaderboard, https://sites.ualberta.ca/~rabelo/COLIEE2023/task1_results.html, accessed on 1 May 2024). SAILER improves retrieval by focusing on the key parts of legal cases: facts, reasoning, and decisions. The model employs a deep encoder model to process the facts of a case and then applies shallow decoders for the reasoning and decision sections of a case law document, which are aggressively masked to force the model to pay closer attention to the relevant information. In the case of statutory law, Long-form Legal Question Answering LLeQA [18] retrieval is also used as a support task to make answering complex legal questions more accurate and interpretable by including relevant documents.

Legal judgment prediction focuses on using AI to forecast the outcomes of legal cases, providing consistency and predictability in judicial decisions. This area of research seeks to leverage AI to improve the fairness and efficiency of legal proceedings. In the study by Guan et al. [19], the authors developed the LDMLSV model (Labor Dispute Machine Learning based on SHapley additive exPlanations and Voting) to predict outcomes in labor dispute cases. This model integrates multiple machine learning algorithms—Random Forest, Extra Trees, and CatBoost—using a soft voting strategy. Notably, it employs Shapley Additive Explanations (SHAP) to provide interpretable predictions, highlighting the key factors influencing case outcomes. The research showed that the model could accurately predict resolution paths.

Sleeman and Gilhooly [20] explored the implications of differing expert judgments in legal contexts, particularly in labor disputes. Their work, grounded in the broader discussion of "Noise" in human judgment, examined how variations in expert decisions can be captured and analyzed using AI models. This research highlights the challenge of inconsistency in expert judgments and suggests that AI could play a crucial role in reducing variability, thereby enhancing the reliability of legal predictions. Lage-Freitas et al. [21] extended this research to the Brazilian legal system, developing models to predict outcomes across a wide range of civil and criminal cases. Their work argued for the versatility of AI in handling diverse legal scenarios and underscored their potential to improve judicial efficiency in complex legal systems.

Legal question answering involves training AI to provide accurate and interpretable answers to legal queries. This area of research focuses on bridging the gap between complex legal information and the general public by offering automated solutions that can deliver expertlike guidance. One notable contribution in this field is the Lawyer LLaMA model [9], which seeks to address the limitations of existing LLMs in handling legal questions. Lawyer LLaMA employs a "retrieve-then-read" pipeline, which first retrieves relevant legal documents and then generates comprehensive answers based on these documents. The model is fine-tuned using a dataset of legal questions, focusing on statutory law. Its design emphasizes the need for interpretability, ensuring that the generated answers are not only accurate but also accompanied by clear rationales that users can verify against the provided legal texts. Similarly, the LLeQA (Long-form Legal Question Answering) [18] system represents a significant advancement in this area. LLeQA is designed to generate long-form answers to legal questions by utilizing a large, expert-annotated dataset. The system employs a retrieval-augmented approach, where a lightweight bi-encoder model retrieves relevant statutory articles, which are then used by an instruction-tuned LLM to generate detailed and interpretable answers. The LLeQA dataset, which includes over 1800 legal questions paired with comprehensive answers, serves as a valuable resource for training and evaluating LLMs in legal question-answering tasks. Despite its promising results on automatic evaluation metrics, the study acknowledges the challenges in ensuring the factual accuracy and relevance of the generated answers, highlighting the ongoing need for refinement in this domain. Both Lawyer LLaMA and LLeQA exemplify the potential of LLMs to democratize access to legal information by providing detailed and reliable answers to complex legal queries.

Document drafting is another important legal task where LLMs are making significant contributions, particularly in the creation of legal documents such as contracts and reports. Weller et al. [22] conducted an in-depth study focused on evaluating and improving LLMs' ability to follow instructions in document drafting tasks. Their research introduced the FOLLOWIR dataset, which was designed to measure how well LLMs could understand and execute detailed instructions during the drafting process. The study highlighted that while modern LLMs achieved impressive capabilities in handling complex instructions across various tasks, they often struggled with tasks that required a deep understanding of specific, nuanced instructions—such as those needed in legal document drafting. The researchers developed FOLLOWIR-7B, a model fine-tuned to better adhere to detailed instructions by leveraging a rigorous evaluation benchmark. This model was shown to significantly improve the precision and relevance of the documents generated, particularly in legal contexts where following exact guidelines is paramount.

Semantic annotation of legal texts uses LLMs to label parts of legal documents, such as identifying rhetorical roles or clause types. This task is very useful to efficiently analyze large volumes of legal text. In their paper, Savelka and Ashley [23] argued that LLMs like GPT-4 could perform these tasks without extensive training, known as zero-shot learning. The study showed that GPT-4 significantly outperformed earlier models in annotating adjudicatory opinions, contractual clauses, and statutory provisions. A key benefit of using LLMs is their capability to function effectively with minimal input, reducing the need for manual annotation. This capability can automate high-volume legal workflows, making complex legal analysis more accessible and cost-effective.

4. Discussion

In this section, we analyze the selected research, focusing on how the latest legal LLMs address the proposed questions. We compare these models based on their characteristics in handling legal matters and their underlying architectures.

The application of LLMs in the legal domain, much like their success in code optimization, holds the potential to significantly transform legal practice, particularly in the drafting of legal documents. Just as developers rely on LLMs to generate and refine code based on specific input parameters [4], lawyers can use these models to create initial drafts of legal documents, such as case reports and contracts [8]. When a lawyer provides the foundational facts of a legal matter—such as the involved parties, relevant laws, and evidence—the AI system can produce a draft document that aligns with legal norms, much like how an LLM processes input to generate functional code. This draft, akin to a preliminary code version, serves as a starting point that the lawyer can then review and refine to ensure it meets the necessary standards of quality and compliance. Additionally, similar to how LLMs can organize and format code into various structures, these models can assist lawyers in organizing legal text into specific formats. While LLMs offer considerable potential in streamlining the creation and formatting of legal documents, it remains essential for lawyers to meticulously review and edit AI-generated drafts to ensure accuracy and adherence to legal standards, just as developers rigorously test and debug code before final deployment.

A good example of a typical emerging pattern in legal AI orchestrations is "In-context learning". This pattern allows leveraging pre-trained LLMs without fine-tuning by controlling their behavior through prompting and conditioning on relevant private data [24]. The key steps involve preprocessing and embedding private legal data into a vector database, constructing prompts by combining templates and examples, and retrieving relevant data from the database based on user queries, executing prompts on LLMs with optional operational systems.

In summary, by applying the concepts from programming to the legal field, LLMs hold the potential to bring significant innovation to many [24] legal practices, making them more efficient, accurate, and accessible. This involves not just automating tasks but also enhancing the cognitive aspects of legal work, such as decision-making, strategy formulation, and personalized legal services.

Beyond drafting, LLMs offer several other significant advantages in legal practice. They can provide automated suggestions for improvement—similar to how LLMs suggest code optimizations [4]. These models can analyze legal documents and propose enhancements based on legal precedents, clarity, and adherence to legal standards. Moreover, LLMs can generate document templates tailored to specific cases or legal requirements, ensuring consistency and compliance with current laws. They are also capable of detecting and correcting errors in legal documents, minimizing risks, and improving the quality of legal filings.

In optimizing legal research, LLMs have proven to be invaluable. They can retrieve precise information from vast legal databases, helping lawyers quickly find relevant precedents, statutes, and case law, thereby streamlining the research process [3]. Additionally, LLMs can perform predictive analysis by examining historical data to forecast outcomes [9] or suggest strategies based on similar cases, aiding lawyers in making informed decisions. Another valuable feature is the ability of LLMs to summarize lengthy legal texts, enabling lawyers to grasp essential details quickly without needing to sift through extensive documents.

LLMs also enhance the efficiency of case management processes. By automating routine tasks, such as scheduling, tracking case progress, and ensuring deadlines are met, these models contribute to overall efficiency [25]. Speech-to-text technology plays a significant role in this area, particularly in court proceedings, where it improves manual record-keeping and enhances transparency [8]. AI-based speech-to-text platforms transcribe spoken language with increasing accuracy, allowing judges, lawyers, and researchers to review and analyze proceedings more effectively [26]. In some international courts, the visible speech-to-text conversion process ensures credibility and increases public trust. All participants benefit from quick access to transcripts, enabling them to identify key arguments and prepare their cases more efficiently. As AI technology continues to evolve, its integration into the judiciary is expected to expand, providing valuable support for decision-making and ensuring the integrity of the legal process. Additionally, LLMs are improving client interaction by automating and enhancing communications through bots or automated responses. This allows for timely updates while freeing up human resources for more complex tasks [18]. Furthermore, data-driven decision-making is being enhanced by leveraging historical case data, which informs strategic decisions, optimizes resource allocation, and predicts case outcomes more accurately [9].

LLMs are continuously learning from historical data, much like how models in programming learn from code edits to suggest performance improvements [4]. In the legal domain, LLMs can learn from new cases [9], legislative changes, and legal publications, improving their suggestions and operations over time. This continuous learning enables the customization and personalization of legal advice and documents, tailoring them to the specific needs and circumstances of each case based on historical outcomes and client preferences [3].

We address all research questions by considering the open-access literature selected in this review and presented in Appendix A, as well as other external resources that back up and further detail the main findings.

RQ1: Which LLM tools are considered leading in the field, and which are best suited for legal applications according to current open-access state-of-the-art research?

The advancement of LLMs in the legal domain has been significantly influenced by the adaptation of open-source foundation models, such as LLaMA, through supervised fine-tuning (SFT) on specialized legal datasets. Table 2 presents a detailed comparison of several Legal LLM models, illustrating their underlying architectures and key features.

Table 2. Legal LLM models.

LLM	Foundation Model	# Params	RLHF	W/API	Origin
LexiLaw [27]	ChatGLM [28]	6B	N	W	CN
Fuzi.mingcha [29,30]	ChatGLM [28]	6B	N	W	CN
LaWGPT-7B-beta1.1	Chinese LLaMA [31]	7B	N	W	CN
Lawyer LLaMA [32]	Chinese LLaMA [31]	13B	N	W	CN
JurisLMs [33]	GPT2 [34], Chinese LLaMA [31]	0.77B/13B	N	W	CN
HanFei [35]	BLOOMZ-7B1 [36]	7B	N	W	CN
ChatLaw [10]	Ziya-LLaMA-13B [37]	13B	N	W	CN
Lychee [38]	GLM-10B [39]	10B	N	W	CN
LLeQA [18]	vicuna-7b-v1.3 [40], wizardLM-7B [41], tulu-7B [42], guanaco-7B [43]	7B	Y	W	EU
FollowIR-7B [22]	Mistral 7B [44]	7B	N	W	US
Lawpath AI [45]	GPT-3.5-turbo-16k [46]	175B	Y	API	AU

ChatLaw [10] is a novel legal LLM that integrates LLMs with vector knowledge databases to mitigate hallucination. It employs robust handling strategies tailored for legal domain challenges. Notably, when evaluated on legal multiple-choice questions using an ELO ranking mechanism [10], ChatLaw outperformed models like Lawyer-Llama [9], LawGPT [47], and GPT-4, demonstrating higher performance in this legal task.

RQ2: What are the primary sources for data extraction and the cutting-edge strategies for dataset development within the legal sector?

Developing robust legal LLMs relies heavily on the quality and diversity of the datasets used during training and fine-tuning. These datasets are essential for equipping models with the necessary knowledge to perform various legal tasks effectively.

Table 3 details the various datasets used in the selected research, extracted from legal corpora, both general and specialized, across multiple jurisdictions. The table highlights how these datasets are employed in different stages of model development, from pretraining to supervised fine-tuning (SFT) and evaluation. This overview includes sources ranging from large-scale raw text corpora to expert-annotated question-answer pairs, demonstrating the diverse approaches taken to enhance the performance of LLMs in the legal domain.

One of the first modern legal datasets [48] is the ECHR dataset consisting of 42 human-annotated judgments from the European Court of Human Rights (https://hudoc.echr.coe.int/eng, accessed on 1 May 2024) [49]. The latest version is annotated for argument mining with three types of argument constituents: premises, conclusions, and non-argument parts. The corpus includes a total of 1951 premises and 743 conclusions.

Similarly, CAIL2018 [50] comprises 2.6 million judgments pertaining to criminal cases published by the Supreme People's Court of China. Each judgment contains a section describing the case's facts, which has been used to predict relevant law articles, charges, and prison terms [50,51].

Another important family of legal Chinese datasets was inspired by the National Judicial Examination of China, a closed-book test required for certification as a lawyer or judge. The exam covers both knowledge-driven questions and questions that require reasoning over aspects from actual case scenarios. The JEC-QA dataset [52] contains 26,365 multiple-choice questions derived from both the national exam as well as practice exercises from other external sources. In a subsequent effort [9], part of the JEC-QA data was augmented with answer explanations to help fine-tune LLMs in the legal domain. The JE-Q2EA and JE-QA2E datasets comprise question–answer pairs and synthetic explanations produced by the GPT series. Since some of the explanations were found to be erroneous, the JE-EXPERT corpus was created using explanations written by law experts.

Beyond fine-tuning for specific tasks, LLMs may benefit from further pretraining on diverse corpora related to the legal domain, such as judicial documents, legal articles, court news, or articles for law popularization [9]. These domain-specific samples may be paired with data from the general domain [53–55] to reduce the risk of catastrophic forgetting of general knowledge, associated with adapting a language model to new tasks [56].

Despite the effort to inject legal knowledge while pretraining a model, it may happen that a relevant legal article does not appear in the pretraining data, or it has changed over time. Thus, it may be helpful to leverage an IR component to augment the input with up-to-date legal articles [9].

Table 3. Data Sources.

Dataset	Application	#Samples
ECHR [49]	Argument detection [48]	1.9k
CAIL2018 [50]	Judgment prediction	2.600k
JEC-QA [51]	Multiple-choice QA	26k
JE-Q2EA [9]	Long-form QA	42k
JE-QA2E [9]	Long-form QA	6k
JE-EXPERT [9]	Long-form QA	850
Legal consultation dataset [57]	Long-form QA [9]	16k
LLeQA [18]	Long-form QA	1.8k
BSARD [58]	Article prediction	1.1k

RQ3: What are the challenges of LLMs in approaching legal tasks?

Legal LLMs, despite their potential, face several significant challenges that complicate their application in legal tasks. One of the primary obstacles is the cost and time associated with inference. As the length of the prompt increases, the computational resources required for processing scale quadratically, and even linear scaling—considered the best theoretical outcome—would be cost-prohibitive for many applications [24]. This challenge has prompted researchers to explore innovative solutions that enhance the reliability and quality of LLM outputs without necessitating a significant increase in the context window [18].

Another critical challenge is the evaluation of truthfulness in LLM outputs. LLMs are prone to generating untruthful or hallucinated content [59], particularly in open-ended text generation scenarios, where distinguishing between accurate and fabricated information can be difficult. To address this issue, a method has been proposed that allows LLMs to interact with external tools to verify their own outputs [60]. This approach draws inspiration from the fact-checking process in journalism, where human-made claims are assessed for truthfulness. By validating the generated content against reliable external sources, the goal is to enhance the truthfulness of LLM outputs [59].

Furthermore, regulating legal LLMs presents unique challenges that require a tailored approach [61]. Current AI regulations are primarily designed for conventional models and do not adequately address the specific needs of legal LLMs. A comprehensive regulatory strategy is necessary [61], encompassing direct regulation, data protection, content moderation, and policy proposals specific to legal LLMs. This approach necessitates the creation of a novel terminology to define the legal LLM value chain, differentiating between developers, deployers, professional/non-professional users, and output recipients. Clear roles and responsibilities [61] along this value chain will enable tailored regulatory duties, fostering the trustworthy development and deployment of legal LLMs for societal benefit.

RQ4: What are the main strategies for increasing the performance of LLMs in addressing legal tasks?

A key strategy to enhance the performance of LLMs in legal tasks is incorporating a Retrieval-Augmented Generation (RAG) module [24]. This approach directly addresses one of the significant challenges LLMs face—hallucinations [59], where the model generates content not grounded in factual information. Hallucinations can be particularly problematic in legal contexts, where accuracy and adherence to legal facts are paramount. The RAG module decreases the error margin by integrating a retrieval mechanism that allows the model to access and incorporate relevant legal documents and sources during the generation process. This retrieval component functions by querying an external knowledge base or database of legal texts, ensuring that the information used to generate responses is accurate and contextually relevant [24]. As a result, the model's outputs become more reliable, as they are anchored in actual legal documents rather than relying solely on the model's internal knowledge, which may be incomplete or outdated. Implementing RAG in LLM applications is emerging as a robust architecture, particularly in domains where faithfulness [59] is critical. According to discussions on evolving LLM architectures, such as those found on [24]'s platform, the RAG framework is not just about improving the reliability of generated content but also about enabling LLMs to scale their capabilities by continuously accessing up-to-date information from large, curated databases. This is particularly beneficial in the legal domain, where the body of knowledge is vast, continuously evolving, and highly detailed.

By deploying a RAG module, LLMs can significantly mitigate the risk of producing unreliable content by decreasing the potential hallucination effect, which could otherwise lead to erroneous or misleading legal advice [9]. This enhancement is crucial for maintaining the trustworthiness and utility of LLMs in legal practice, ensuring that their outputs are not only increasingly more accurate but also more grounded in the most relevant and current legal texts available. Consequently, RAG modules are becoming a critical component in the architecture of LLMs designed for legal applications, reflecting a broader trend toward integrating retrieval-based techniques to enhance the performance and reliability of AI systems across various high-stakes domains.

Another strategy to enhance LLM performance in specialized legal tasks has been the generation of synthetic data for supervised fine-tuning (SFT) by distilling knowledge from powerful LLMs like the GPT series. Various studies have explored this approach, including the works of [62,63]. Interestingly, some researchers have observed that in specific instances, data generated by the GPT series can exhibit greater diversity and usefulness compared to human-written data, particularly for tasks in general domains, as noted by [9,62,64]. However, in a specialized field such as the legal domain, fine-tuning on a small collection of expert-written training samples may still produce better results than leveraging a much larger synthetic dataset [9].

Specialized model training for different legal scenarios is also gaining traction. Rather than relying on a single general-purpose legal LLM, distinct models could be developed for tasks such as multiple-choice questions, keyword extraction, and question answering. To manage the selection and deployment of these specialized models, a large LLM was employed as a controller in [10]. This controller model dynamically determined which was most suitable for a given user request, ensuring that the appropriate model was utilized for the task at hand. By tailoring models to specific legal tasks and leveraging a controller LLM to orchestrate their deployment, this approach aimed to optimize performance and provide more accurate and relevant outputs for various legal scenarios rather than relying on a one-size-fits-all solution.

AI agentic workflows, as presented by Ng [65], are applied to code generation in his examples, but we can easily imagine a very similar approach applied to legal tasks. The reflection design pattern involves prompting an LLM to iteratively improve its output through self-criticism and feedback rather than generating a final output directly. This process enables the LLM to identify potential issues in its initial response and provide constructive suggestions for enhancement. Reflection can be augmented by providing tools for output evaluation, like unit tests or web searches. Additionally, it can be implemented using a multi-agent framework [60] with separate agents for output generation and constructive criticism, leading to a discussion that improves responses. Overall, reflection leverages an LLM's ability for self-evaluation and iterative refinement, potentially driving significant performance gains across various tasks [65].

Tool use enables LLMs to validate and iteratively refine their outputs using external tools, similar to how humans utilize tools for cross-checking and improving their work. For example, by interacting with appropriate tools to evaluate aspects of the generated text, and then revising based on the feedback, CRITIC [60] consistently enhances LLM performance across various tasks. This pattern, similar to reflection from this perspective, underscores the importance of external feedback loops in promoting the self-improvement capabilities of LLMs [65].

Planning: LLMs have shown noteworthy intelligence, inducing a growing interest in leveraging them as planning modules for autonomous agents tasked with perceiving environments and executing actions. Planning, a critical capability requiring complex reasoning and decision-making, could benefit from LLMs' impressive performance across domains like reasoning, tool usage, and instruction-following [66]. Consequently, various methodologies aim to harness LLMs' intelligence as cognitive cores to enhance autonomous agents' planning abilities. To systematically explore these methodologies, ref. [66] proposes a novel taxonomy that categorizes existing works on LLM-based agent planning into five key areas: task decomposition, multi-plan selection, external module-aided planning, reflection and refinement, and memory-augmented planning. This taxonomy provides a structured framework to analyze and compare approaches to integrating legal LLMs into autonomous agent architectures for improved planning capabilities.

Multi-agent collaboration: Agent-planning methodologies can be crucial in optimizing multi-agent collaboration to enhance the factual accuracy and reliability of LLM outputs [67]. By incorporating task decomposition approaches, the complex legal reasoning process can be broken down into more manageable sub-tasks distributed across multiple specialized LLM agents. These agents can then collaborate, exchanging intermediate outputs and leveraging external knowledge sources or modules to verify and refine their results iteratively [60,66]. Reflection and refinement techniques enable the agents to critically evaluate their outputs, identify potential inaccuracies or gaps, and engage in a cyclic process of improvement, drawing upon the collective intelligence of the multi-agent system [65]. Memory-augmented planning can further enhance this process by allowing agents to maintain and reference relevant contextual information, legal precedents, or domain-specific knowledge throughout the collaborative reasoning process [66].

Ultimately, this multi-agent approach, guided by agent-planning methodologies, can lead to a more robust and trustworthy legal analysis by mitigating the hallucination tendencies of individual LLMs [59] and enabling an iterative refinement process within the agent collective.

RQ5: What are the main limitations of current LLMs for the legal domain?

Several limitations and shortcomings of existing approaches have been identified in the reviewed literature.

The overconfidence when providing inaccurate data is one of the biggest limitations. This can be noticed in all the SOTA foundational LLM models like GPT-4, Claude 3, Gemini, and LLama-2/3. LLMs tend to generate fluent but inaccurate texts, a phenomenon known as hallucination [68]. For individuals without specialized knowledge, it can be challenging to identify such unfaithful outputs, which may lead to misinformation and potential harm, especially in the legal domain.

For example, as legal articles serve as crucial evidence in the legal field, the researchers behind Lawyer LLaMA [9] focus on two types of hallucinations related to these articles:

- Whether the model fabricates one or more nonexistent legal articles.
- If the response mentions an existing legal article, whether it incorrectly quotes the title of the law or the article's number.

By assessing the model's performance in these two areas, researchers can determine the effectiveness of the retrieval module in mitigating hallucinations and enhancing the reliability of the model's outputs when dealing with legal articles, which are of utmost importance in the legal domain.

The multilingual and multicultural nature of the law poses significant challenges for applying big data and AI in the legal domain. As legal systems span various languages and cultures, legal big data often comprise texts in different languages, requiring a cross-lingual analysis [8].

One of the primary challenges in dealing with multilingual legal data is translation. The nature of the text to be translated plays a central role in legal translation, as legal texts require not only linguistic but also contextual and cultural precision [69]. Legal terminology and concepts can vary significantly across languages and countries, making accurate translation crucial for understanding and applying legal principles. Differences in cultural norms and legal traditions further increase the difficulty in interpreting legal texts across languages [70]. For instance, European Union regulations are often published in multiple official language versions, such as English, French, or German. The European Union serves as a best-practice example (https://www.europarl.europa.eu/RegData/etudes/BRIE/2017/595914/EPRS_BRI(2017)595914_EN.pdf, accessed on 29 September 2024), where legal translations are effectively managed through a hybrid system combining custom AI tools (https://language-tools.ec.europa.eu/, accessed on 29 September 2024; eTranslation—https://commission.europa.eu/resources-partners/etranslation_en, accessed on 29 September 2024) with specialized legal translation experts. This ensures the accurate transfer of legal meanings across languages while minimizing misinterpretation risks. Additionally, the EU's multilingual regulation dataset could potentially serve as a highly valuable resource for training and developing legal AI systems. Its rigorously translated legal texts across multiple languages offer an ideal foundation for advancing multilingual applications in the legal domain. However, even with this robust system, legal researchers and practitioners must still carefully compare and analyze the different language versions to ensure the precise understanding of the regulations' meanings and implications [70], as inconsistencies can lead to legal ambiguities [69].

Moreover, while translation technologies and AI models have made great strides, many foundational models, including the GPT series [1], LLaMA series [71], and similar architectures, were primarily trained on large datasets from the Internet with a relatively low proportion of legal texts. As a result, these models often lack the depth of understanding required to handle the nuances and subtleties of legal language and semantic translation. For instance, legal phrases that include modal verbs such as "shall" or "may" often carry specific and nuanced legal implications [72], which are not consistently captured by current LLMs. These verbs can denote obligations or permissions, and their correct interpretation is vital in legal contexts. Similarly, expressions of "necessity" or "permission", often conveyed through modal verbs like "must" or "may", are critical for legal interpretation. However, AI systems frequently struggle to translate these concepts with full semantic accuracy.

Additionally, legal texts often depend on precise logical constructs, such as the operators "and", "or", and "not", which are fundamental in classical logic. When these logical operators are used in conjunction with modal verbs, the accuracy of interpretation becomes even more crucial. Any failure by LLMs to accurately interpret both the modal meanings [72] and the logical operators within a legal text could lead to significant errors, causing further misinterpretations impacting the overall trustworthiness of the AI system. These combined challenges in handling both modality and logic contribute to potential inconsistencies in the AI's output.

Legal AI systems should consider cultural differences and variations in legal terminology when analyzing multilingual data. This consideration may necessitate collaboration among legal experts, linguists, and AI researchers to develop culturally sensitive, context-aware models that more accurately interpret and apply legal principles across different languages and jurisdictions. However, although there is a substantial quantity of legal data available for training, it is relatively small compared to the vast corpus of general web data used to pretrain LLMs. This imbalance indicates that legal content constitutes only a minor fraction of their overall knowledge base, potentially impacting their performance on legal tasks. Furthermore, LLMs may struggle with context fidelity, often relying on their extensive parametric knowledge acquired during training [73] rather than focusing exclusively on the specific text at hand. When pre-existing knowledge conflicts with the input text, this can lead to inaccuracies. In legal contexts, where precise interpretation and strict adherence to the provided text are crucial, such tendencies might result in significant errors and misinterpretations.

The vast scale and complexity of legal data pose significant challenges for AI systems in the legal domain. Legal datasets often contain lengthy documents, such as legislation and court judgments, which require powerful computing capabilities and efficient algorithms to process and analyze effectively [74].

Moreover, legal data vary greatly across domains, each with unique regulations and document types. This diversity necessitates specialized analysis methods and the incorporation of domain-specific knowledge into AI systems. However, LLMs face challenges related to data retention, as they cannot selectively delete specific pieces of information from their training data. This inability poses problems when certain data become obsolete or need to be removed due to legal requirements [75]. Compliance with privacy laws, such as the EU General Data Protection Regulation (GDPR), is crucial, as these regulations grant individuals the "right to be forgotten". Similarly, tax law is highly complex, with extensive regulations and judicial precedents [8]. To address these challenges, AI systems must leverage robust computational power, advanced algorithms, and domain-specific knowledge. Close collaboration between legal experts and AI researchers is crucial for developing AI models that align with legal principles and practices.

To sum up, the vast scale and complexity of legal data require powerful computing capabilities, specialized analysis methods, and the incorporation of domain-specific knowledge [8]. By fostering interdisciplinary collaboration and leveraging advanced computational techniques, AI systems can effectively navigate the complex landscape of legal data and provide valuable insights to support legal professionals.

The ever-changing nature of the law poses a significant challenge for legal AI systems. Legal documents and regulations are frequently revised, necessitating regular updates of legal datasets to reflect the current provisions. However, once an LLM is trained, its knowledge base becomes static. It cannot inherently update itself with new information or remove outdated data without retraining, making its inability to delete obsolete data a significant limitation. This characteristic hampers the model's capability to stay current with legal developments and may lead to the dissemination of outdated or incorrect legal information. This is particularly important in fields like tax law, where new regulations are introduced to adapt to changing economic conditions. To address this challenge, legal big data systems must be designed with the flexibility to accommodate frequent updates and seamlessly incorporate new legal information. AI models must also be able to adapt to these changes, provide up-to-date analysis based on the most current legal provisions, and contextualize to a required time interval. For example, to assess the litigation result based on the legislation applicable on the date that the contract was signed.

However, some foundation LLMs (LLMs) used in legal AI systems are trained on data up to a certain cut-off date and lack access to real-time information from the internet. Their knowledge can become obsolete over time, as they cannot incorporate the latest legal changes and updates. To mitigate this issue, ongoing cooperation between legal experts and AI researchers is crucial to ensure that legal AI systems remain aligned with the latest legal developments. Legal professionals can guide the updating process and provide insights into the implications of legal changes, helping to maintain the accuracy and relevance of AI-generated analysis [8].

As such, the ever-evolving nature of the law requires legal AI systems to be designed with flexibility and adaptability. However, the static nature of some LLMs and foundation models can lead to obsolescence, requiring a long-term partnership between legal specialists and AI developers in order to maintain the accuracy and relevance of legal AI systems. We must emphasize that **privacy and security are paramount when dealing with legal systems**. Legal documents often contain sensitive information, such as personal identifying details, which must be strictly protected to maintain confidentiality and prevent unauthorized access [8].

Anonymization techniques, such as removing or obscuring personal information, should be applied during the collection, storage, and analysis of legal data. However, anonymization alone is not sufficient. Legal big data systems must incorporate robust security measures, including encryption, access control, and secure storage protocols, to safeguard sensitive information from breaches and malicious activities. Designing legal AI systems with privacy and security in mind is crucial. Techniques like differential privacy can enable the analysis of legal data while preserving individual privacy. Regular auditing and monitoring of these systems are necessary to detect and prevent potential privacy or security breaches.

From an IT perspective, implementing strong access control mechanisms, such as role-based access control (RBAC) and multi-factor authentication (MFA), can help ensure that only authorized individuals can access sensitive legal data. Encryption techniques, like end-to-end encryption and homomorphic encryption, can protect data both in transit and at rest. Moreover, homomorphic encryption can enable secure outsourcing of legal data processing to third-party service providers, such as cloud computing platforms. Law firms and legal organizations can encrypt their sensitive data before uploading them to the cloud, ensuring that the service provider cannot access the plaintext data. The service provider can still perform computations and analysis on the encrypted data, such as searching for specific legal precedents or running machine learning algorithms, without decrypting the data. This allows legal organizations to leverage the scalability and computational power of cloud computing while maintaining the confidentiality of their data [76].

Moreover, establishing clear data governance policies and adhering to relevant privacy regulations, such as the General Data Protection Regulation (GDPR) or the Health Insurance Portability and Accountability Act (HIPAA), is essential to ensure compliance and maintain the trust of individuals whose data are being processed.

In summary, we must emphasize the critical importance of privacy and security in legal big data systems. Robust technical measures, including anonymization, encryption, and access control, must be implemented alongside regular auditing and monitoring. Designing legal AI systems with privacy-preserving techniques and establishing strong data governance policies are crucial to maintaining the confidentiality and integrity of sensitive legal information.

Limitations

This review offers valuable insights but also faces limitations. Some studies show inconsistencies across legal subdomains and geographic focus. Additionally, the rapid evolution of AI means the findings may soon become outdated. The review reflects research as of 28 March 2024, but delays between study completion and publication mean some issues may have already been tackled in newer legal LLM versions. Furthermore, due to resource constraints, the screening and data extraction were conducted by a single reviewer, which may have introduced the risk of bias and errors in study selection and data interpretation. To mitigate this limitation, a second, more experienced reviewer was consulted to validate coding decisions. However, the absence of multiple independent reviewers throughout the entire process may still affect the reliability of the results.

5. Conclusions

A significant gap remains between the potential of these technologies and their practical application in legal settings. These specialized language models, despite their potential, often struggle to capture the multifaceted complexity of legal systems.

Integrated approaches, which combine LLMs with vector knowledge databases [9], use in-context learning [24] and/or employ multi-agent architectures [60,66], exhibited enhanced performance on legal tasks compared to standalone LLMs like GPT-3 and Lawyer-LLaMA. By leveraging external legal knowledge sources, iterative refinement through agents [65], and robust handling strategies, these integrated systems achieve competitive or superior performance to human experts on multiple-choice legal questions [9] while mitigating hallucination tendencies to a significant degree.

However, the performance of LLMs varies across different legal domains and task types, with more complex tasks like legal writing and reasoning posing challenges due to LLMs' tendencies for hallucination [59] and lack of robust legal knowledge grounding. Sometimes the conclusions are contradictory, e.g., Ng [65] argued for a significant improvement of performance using agentic reflection by significantly surpassing GPT 4.0's shot performance with GPT 3.5 using few-shot reflection; in contrast, Gou et al. [60] found that relying solely on self-correction without incorporating external feedback may result in only modest improvements or could even lead to a decline in performance.

Integrating LLMs with external legal knowledge sources, vector databases, and retrieval mechanisms can significantly enhance their accuracy [24], domain coverage, and trustworthiness, mitigating issues like hallucination while leveraging their language understanding capabilities. Observing the current rate of improvement in LLMs, it is reasonable to anticipate that the performance of foundational LLMs will continue to improve significantly in the future. To address the current limitations and maximize future benefits, applications must be designed with upgradable pipelines that can quickly integrate newer versions of foundational LLMs, ensuring that they remain at the cutting edge of legal AI capabilities.

Author Contributions: Conceptualization, B.P. and M.D.; methodology, B.P. and M.D.; validation, B.P., R.I., T.R. and M.D.; investigation, B.P. and M.D.; resources, B.P., R.I. and M.D.; data curation, B.P. and M.D.; writing—original draft preparation, B.P. and R.I.; writing—review and editing, T.R. and M.D.; supervision, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The material used in this study is available from the sourced databases (accessed on 28 March 2024).

Acknowledgments: PRISMA2020 Flow Diagram https://github.com/prisma-flowdiagram/PRISMA2020, accessed on 1 March 2024.

Conflicts of Interest: Author Traian Rebedea was employed by the company NVIDIA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AAAI	Association for the Advancement of Artificial Intelligence
AI	Artificial intelligence
Arxiv	An open-access repository of preprints
ChatGPT	Chat Generative Pre-trained Transformer
COLIEE	Competition on Legal Information Extraction and Entailment

DOAJ	Directory of open access journals
ECHR	European Court of Human Rights
GPT	Generative Pre-trained Transformer
ICAIL	The International Conference on Artificial Intelligence and Law
JURIX	The International Conference on Legal Knowledge and Information Systems
JURISIN	The International Workshop on Juris-informatics
JE-Q2EA	Judicial Examination-Question to Explanation + Answer
JE-QA2E	Judicial Examination-Question + Answer to Explanation
JE-EXPERT	Judicial Examination-Expert Corpus
LLM	Large Language Model
MDPI	Multidisciplinary Digital Publishing Institute
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback
SCL	Society for Computers and Law
SOTA	State of the art
SFT	Supervised fine-tuning

Appendix A. Selected Articles

Title	Authors	URL (accessed on 28 March 2024)	Year
1. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models	Louis, Antoine; van Dijck, Gijs; Spanakis, Gerasimos	http://arxiv.org/abs/2309.17050	2023
2. Lawyer LLaMA Technical Report	Huang, Quzhe; Tao, Mingxu; Zhang, Chen; An, Zhenwei; Jiang, Cong; Chen, Zhibin; Wu, Zirui; Feng, Yansong	http://arxiv.org/abs/2305.15062	2023
3. Instruction Tuning with GPT-4	Peng, Baolin; Li, Chunyuan; He, Pengcheng; Galley, Michel; Gao, Jianfeng	http://arxiv.org/abs/2304.03277	2023
4. Predicting Brazilian Court Decisions	Lage-Freitas, André; Allende-Cid, Héctor; Santana, Orivaldo; Oliveira-Lage, Lívia	https://peerj.com/articles/cs-904	2022
5. Hammering with the Telescope	Sobkowicz, Pawel	https://www.frontiersin.org/articles/10.3389/frai.2022.1010219/full	2022
6. GiusBERTo: Italy's AI-Based Judicial Transformation: A Teaching Case	Datta, Pratim; Zahn, Brian J.; Attias, Luca; Salierno, Giulio; Bertè, Rosamaria; Battisti, Daniela; Acton, Thomas	https://aisel.aisnet.org/cais/vol53/iss1/33/	2023
7. Regulating ChatGPT and other Large Generative AI Models	Hacker, Philipp; Engel, Andreas; Mauer, Marco	https://dl.acm.org/doi/10.1145/3593013.3594067	2023
8. Prediction Machine Learning Models on Propensity Convicts to Criminal Recidivism	Kovalchuk, Olha; Karpinski, Mikolaj; Banakh, Serhiy; Kasianchuk, Mykhailo; Shevchuk, Ruslan; Zagorodna, Nataliya	https://www.mdpi.com/2078-2489/14/3/161	2023
9. Machine Learning in Bail Decisions and Judges' Trustworthiness	Morin-Martel, Alexis	https://link.springer.com/10.1007/s00146-023-01673-6	2023
10. Regression Applied to Legal Judgments to Predict Compensation for Immaterial Damage	Dal Pont, Thiago Raulino; Sabo, Isabela Cristina; Hübner, Jomi Fred; Rover, Aires José	https://peerj.com/articles/cs-1225	2023
11. How To Build The Ultimate Legal LLM Stack	Dominic Woolrych	https://www.linkedin.com/pulse/how-build-ultimate-legal-llm-stack-dominic-woolrych/	2023
12. Emerging Architectures for LLM Applications	Matt Bornstein, Rajko Radovanovic	https://a16z.com/emerging-architectures-for-llm-applications/?trk=article-ssr-frontend-pulse_little-text-block	2023

Title	Authors	URL (accessed on 28 March 2024)	Year
13. LegalVis: Exploring and Inferring Precedent Citations in Legal Documents	Resck, Lucas E.; Ponciano, Jean R.; Nonato, Luis Gustavo; Poco, Jorge	https://ieeexplore.ieee.org/document/9716779/	2023
14. Emerging Trends: Smooth-talking Machines	Church, Kenneth Ward; Yue, Richard	https://www.cambridge.org/core/product/identifier/S1351324923000463/type/journal_article	2023
15. The Unreasonable Effectiveness of Large Language Models in Zero-shot Semantic Annotation of Legal Texts	Savelka, Jaromir; Ashley, Kevin D.	https://www.frontiersin.org/articles/10.3389/frai.2023.1279794/full	2023
16. Groups of Experts Often Differ in Their Decisions: What are the Implications for AI and Machine Learning?	Sleeman, Derek H.; Gilhooly, Ken	https://onlinelibrary.wiley.com/doi/10.1002/aaai.12135	2023
17. Predicting Critical Path of Labor Dispute Resolution in Legal Domain by Machine Learning Models Based on SHapley Additive ExPlanations and Soft Voting Strategy	Guan, Jianhua; Yu, Zuguo; Liao, Yongan; Tang, Runbin; Duan, Ming; Han, Guosheng	https://www.mdpi.com/2227-7390/12/2/272	2024
18. THUIR@COLIEE 2023: Incorporating Structural Knowledge into Pre-trained Language Models for Legal Case Retrieval	Li, Haitao; Su, Weihang; Wang, Changyue; Wu, Yueyue; Ai, Qingyao; Liu, Yiqun	http://arxiv.org/abs/2305.06812	2023
19. The Implications of ChatGPT for Legal Services and Society	Perlman, Andrew	https://www.ssrn.com/abstract=4294197	2022
20. GPT Takes the Bar Exam	Bommarito, Michael; Katz, Daniel Martin	http://arxiv.org/abs/2212.14402	2022
21. Unlocking Practical Applications in Legal Domain	Savelka, Jaromir	https://dl.acm.org/doi/10.1145/3594536.3595161	2023
22. Large Language Models in Law: A Survey	Lai, Jinqi; Gan, Wensheng; Wu, Jiayang; Qi, Zhenlian; Yu, Philip S.	http://arxiv.org/abs/2312.03718	2023
23. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking	Zelikman, Eric; Harik, Georges; Shao, Yijia; Jayasiri, Varuna; Haber, Nick; Goodman, Noah D.	http://arxiv.org/abs/2403.09629	2024
24. FollowIR: Evaluating and Teaching Information Retrieval Models to Follow Instructions	Weller, Orion; Chang, Benjamin; MacAvaney, Sean; Lo, Kyle; Cohan, Arman; Van Durme, Benjamin; Lawrie, Dawn; Soldaini, Luca	http://arxiv.org/abs/2403.15246	2024
25. Performance Analysis of Large Language Models in the Domain of Legal Argument Mining	Al Zubaer, Abdullah; Granitzer, Michael; Mitrović, Jelena	https://www.frontiersin.org/articles/10.3389/frai.2023.1278796/full	2023
26. A Dynamic Approach for Visualizing and Exploring Concept Hierarchies from Textbooks	Wehnert, Sabine; Chedella, Praneeth; Asche, Jonas; De Luca, Ernesto William	https://www.frontiersin.org/articles/10.3389/frai.2024.1285026/full	2024
27. The Benefits and Dangers of Using Machine Learning to Support Making Legal Predictions	Zeleznikow, John	https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1505	2023
28. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases	Cui, Jiaxi; Li, Zongjian; Yan, Yang; Chen, Bohua; Yuan, Li	http://arxiv.org/abs/2306.16092	2023
29. Survey of Hallucination in Natural Language Generation	Ji, Ziwei; Lee, Nayeon; Frieske, Rita; Yu, Tiezheng; Su, Dan; Xu, Yan; Ishii, Etsuko; Bang, Yejin; Chen, Delong; Chan, Ho Shu; Dai, Wenliang; Madotto, Andrea; Fung, Pascale	http://arxiv.org/abs/2202.03629	2022
30 Long-form Factuality in Large Language Models	Wei, Jerry; Yang, Chengrun; Song, Xinying; Lu, Yifeng; Hu, Nathan; Tran, Dustin; Peng, Daiyi; Liu, Ruibo; Huang, Da; Du, Cosmo; Le, Quoc V.	http://arxiv.org/abs/2403.18802	2024
31. Understanding the Planning of LLM Agents: A Survey	Huang, Xu; Liu, Weiwen; Chen, Xiaolong; Wang, Xingmei; Wang, Hao; Lian, Defu; Wang, Yasheng; Tang, Ruiming; Chen, Enhong	http://arxiv.org/abs/2402.02716	2024

Title	Authors	URL (accessed on 28 March 2024)	Year
32. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models	Guha, Neel; Nyarko, Julian; Ho, Daniel E.; Ré, Christopher; Chilton, Adam; Narayana, Aditya; Chohlas-Wood, Alex; Peters, Austin; Waldon, Brandon; Rockmore, Daniel N.; Zambrano, Diego; Talisman, Dmitry; Hoque, Enam; Surani, Faiz; Fagan, Frank; Sarfaty, Galit; Dickinson, Gregory M.; Porat, Haggai; Hegland, Jason; Wu, Jessica; Nudell, Joe; Niklaus, Joel; Nay, John; Choi, Jonathan H.; Tobia, Kevin; Hagan, Margaret; Ma, Megan; Livermore, Michael; Rasumov-Rahe, Nikon; Holzenberger, Nils; Kolt, Noam; Henderson, Peter; Rehaag, Sean; Goel, Sharad; Gao, Shang; Williams, Spencer; Gandhi, Sunny; Zur, Tom; Iyer, Varun; Li, Zehua	http://arxiv.org/abs/2308.11462	2023
33. LawBench: Benchmarking Legal Knowledge of Large Language Models	Fei, Zhiwei; Shen, Xiaoyu; Zhu, Dawei; Zhou, Fengzhe; Han, Zhuo; Zhang, Songyang; Chen, Kai; Shen, Zongwen; Ge, Jidong	http://arxiv.org/abs/2309.16289	2023

References

1. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
2. Sobkowicz, P. Hammering with the telescope. *Front. Artif. Intell.* **2022**, *5*, 1010219. [CrossRef] [PubMed]
3. Villata, S.; Araszkiewicz, M.; Ashley, K.; Bench-Capon, T.; Branting, L.K.; Conrad, J.G.; Wyner, A. Thirty years of artificial intelligence and law: The third decade. *Artif. Intell. Law* **2022**, *30*, 561–591. [CrossRef]
4. Ridnik, T.; Kredo, D.; Friedman, I. Code Generation with AlphaCodium: From Prompt Engineering to Flow Engineering. *arXiv* **2024**, arXiv:2401.08500.
5. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-Verification Reduces Hallucination in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand, 8 August 2024; pp. 3563–3578.
6. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv* **2022**, arXiv:2203.11171.
7. Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Zhang, S.; Chen, K.; Shen, Z.; Ge, J. LawBench: Benchmarking Legal Knowledge of Large Language Models. *arXiv* **2023**, arXiv:2309.16289.
8. Lai, J.; Gan, W.; Wu, J.; Qi, Z.; Yu, P.S. Large Language Models in Law: A Survey. *arXiv* **2023**, arXiv:2312.03718. [CrossRef]
9. Huang, Q.; Tao, M.; Zhang, C.; An, Z.; Jiang, C.; Chen, Z.; Wu, Z.; Feng, Y. Lawyer LLaMA Technical Report. *arXiv* **2023**, arXiv:2305.15062.
10. Cui, J.; Li, Z.; Yan, Y.; Chen, B.; Yuan, L. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. *arXiv* **2023**, arXiv:2306.16092.
11. Re, R.M.; Solow-Niederman, A. Developing artificially intelligent justice. *Stanf. Technol. Law Rev.* **2019**, *22*, 242.
12. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473.
13. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
14. Haddaway, N.R.; Page, M.J.; Pritchard, C.C.; McGuinness, L.A. *PRISMA2020*: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Syst. Rev.* **2022**, *18*, e1230. [CrossRef] [PubMed]
15. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python; Explosion: 2020. Available online: https://spacy.io (accessed on 28 March 2024).
16. Grootendorst, M. MaartenGr/KeyBERT: BibTeX (Version v0.1.3); Zenodo: 2021. Available online: https://zenodo.org/records/4461265 (accessed on 28 March 2024). [CrossRef]
17. Li, H.; Su, W.; Wang, C.; Wu, Y.; Ai, Q.; Liu, Y. THUIR@COLIEE 2023: Incorporating Structural Knowledge into Pre-trained Language Models for Legal Case Retrieval. *arXiv* **2023**, arXiv:2305.06812.
18. Louis, A.; van Dijck, G.; Spanakis, G. Interpretable long-form legal question answering with retrieval-augmented large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 22266–22275.

19. Guan, J.; Yu, Z.; Liao, Y.; Tang, R.; Duan, M.; Han, G. Predicting Critical Path of Labor Dispute Resolution in Legal Domain by Machine Learning Models Based on SHapley Additive exPlanations and Soft Voting Strategy. *Mathematics* **2024**, *12*, 272. [CrossRef]
20. Sleeman, D.H.; Gilhooly, K. Groups of experts often differ in their decisions: What are the implications for AI and machine learning? A commentary on *Noise: A Flaw in Human Judgment*, by Kahneman, Sibony, and Sunstein (2021). *AI Mag.* **2023**, *44*, 555–567. [CrossRef]
21. Lage-Freitas, A.; Allende-Cid, H.; Santana, O.; Oliveira-Lage, L. Predicting Brazilian Court Decisions. *Peerj Comput. Sci.* **2022**, *8*, e904. [CrossRef]
22. Weller, O.; Chang, B.; MacAvaney, S.; Lo, K.; Cohan, A.; Durme, B.V.; Lawrie, D.; Soldaini, L. FollowIR: Evaluating and Teaching Information Retrieval Models to Follow Instructions. *arXiv* **2024**, arXiv:2403.15246.
23. Savelka, J.; Ashley, K.D. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Front. Artif. Intell.* **2023**, *6*, 1279794. [CrossRef]
24. Bornstein Matt, R.R. Emerging Architectures for LLM Applications. 2023. Available online: https://a16z.com (accessed on 20 June 2023).
25. Xu, Z. Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities. *Appl. Artif. Intell.* **2022**, *36*, 2013652. [CrossRef]
26. Etulle, R.D.; Moslares, F.; Pacad, E.; Odullo, J.; Nacionales, J.; Claridad, N. Investigating the Listening and Transcription Performance in Court: Experiences from Stenographers in Philippine Courtrooms. *J. Lang. Pragmat. Stud.* **2023**, *2*, 100–111. [CrossRef]
27. Haitao, L. LexiLaw. 2023. Available online: https://github.com/CSHaitao/LexiLaw (accessed on 1 May 2024).
28. GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv* **2024**, arXiv:2406.12793.
29. Wu, S.; Liu, Z.; Zhang, Z.; Chen, Z.; Deng, W.; Zhang, W.; Yang, J.; Yao, Z.; Lyu, Y.; Xin, X.; et al. fuzi.mingcha. 2023. Available online: https://github.com/irlab-sdu/fuzi.mingcha (accessed on 28 March 2024).
30. Deng, W.; Pei, J.; Kong, K.; Chen, Z.; Wei, F.; Li, Y.; Ren, Z.; Chen, Z.; Ren, P. Syllogistic Reasoning for Legal Judgment Analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 13997–14009. [CrossRef]
31. Cui, Y.; Yang, Z.; Yao, X. Efficient and effective text encoding for chinese llama and alpaca. *arXiv* **2023**, arXiv:2304.08177.
32. Huang, X.; Zhang, L.L.; Cheng, K.T.; Yang, F.; Yang, M. Fewer is More: Boosting LLM Reasoning with Reinforced Context Pruning. *arXiv* **2023**, arXiv:2312.08901.
33. JurisLMs. 2023. Available online: https://github.com/seudl/JurisLMs (accessed on).
34. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI blog* **2019**, *1*, 9.
35. He, W.; Wen, J.; Zhang, L.; Cheng, H.; Qin, B.; Li, Y.; Jiang, F.; Chen, J.; Wang, B.; Yang, M. HanFei-1.0. 2023. Available online: https://github.com/siat-nlp/HanFei (accessed on 28 March 2024).
36. Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T.L.; Bari, M.S.; Shen, S.; Yong, Z.X.; Schoelkopf, H.; et al. Crosslingual generalization through multitask finetuning. *arXiv* **2022**, arXiv:2211.01786.
37. Zhang, J.; Gan, R.; Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; et al. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *arXiv* **2022**, arXiv:2209.02970.
38. Shen, X.; Zhu, D.; Fei, Z.; Li, Q.; Shen, Z.; Ge, J. Lychee. 2023. Available online: https://github.com/davidpig/lychee_law (accessed on 28 March 2024).
39. Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 320–335. [CrossRef]
40. Chiang, W.L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J.E.; et al. Vicuna: An Open-Source Chatbot Impressing Gpt-4 with 90%* Chatgpt Quality. 2023. Available online: https://lmsys.org/blog/2023-03-30-vicuna (accessed on 28 March 2024).
41. Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. *arXiv* **2023**, arXiv:2304.12244.
42. Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K.R.; Wadden, D.; MacMillan, K.; Smith, N.A.; Beltagy, I.; et al. How far can camels go? In exploring the state of instruction tuning on open resources. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 74764–74786.
43. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv* **2023**, arXiv:2305.14314.
44. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.

45. Woolrych, D. How To Build The Ultimate Legal LLM Stack. 2023. Available online: https://lawpath.com.au/blog/how-to-build-the-ultimate-legal-llm-stack (accessed on 28 March 2024).
46. OpenAI. GPT-3.5-turbo-16k. 2023. Available online: https://openai.com (accessed on 1 May 2024).
47. Nguyen, H.T. A Brief Report on LawGPT 1.0: A Virtual Legal Assistant Based on GPT-3. *arXiv* **2023**, *arXiv:2302.05729*.
48. Moens, M.F.; Boiy, E.; Palau, R.M.; Reed, C. Automatic detection of arguments in legal texts. In Proceedings of the 11th International Conference on Artificial Intelligence and Law, ICAIL '07, New York, NY, USA, 4–8 June 2007; pp. 225–230. [CrossRef]
49. Zubaer, A.A.; Granitzer, M.; Mitrović, J. Performance analysis of large language models in the domain of legal argument mining. *Front. Artif. Intell.* **2023**, *6*, 1278796. [CrossRef] [PubMed]
50. Xiao, C.; Zhong, H.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; et al. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv* **2018**, arXiv:1807.02478.
51. Zhong, H.; Zhou, J.; Qu, W.; Long, Y.; Gu, Y. An Element-aware Multi-representation Model for Law Article Prediction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6663–6668. [CrossRef]
52. Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; Sun, M. JEC-QA: A legal-domain question answering dataset. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 9701–9708.
53. Yuan, S.; Zhao, H.; Du, Z.; Ding, M.; Liu, X.; Cen, Y.; Zou, X.; Yang, Z.; Tang, J. WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. *AI Open* **2021**, *2*, 65–68. [CrossRef]
54. Xu, L.; Zhang, X.; Dong, Q. CLUECorpus2020: A Large-scale Chinese Corpus for Pre-training Language Model. *arXiv* **2020**, arXiv:2003.01355.
55. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
56. Chen, S.; Hou, Y.; Cui, Y.; Che, W.; Liu, T.; Yu, X. Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7870–7881.
57. Chen, F. The Legal Consultation Data and Corpus of the Thesis from China Law Network (Version V1); Peking University Open Research Data Platform. 2018. Available online: https://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/OLO4G8 (accessed on 28 March 2024).
58. Louis, A.; Spanakis, G. A Statutory Article Retrieval Dataset in French. *arXiv* **2022**, arXiv:2108.11792.
59. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
60. Gou, Z.; Shao, Z.; Gong, Y.; Shen, Y.; Yang, Y.; Duan, N.; Chen, W. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. *arXiv* **2023**, arXiv:2305.11738.
61. Hacker, P. The European AI liability directives—Critique of a half-hearted approach and lessons for the future. *Comput. Law Secur. Rev.* **2023**, *51*, 105871. [CrossRef]
62. Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N.A.; Khashabi, D.; Hajishirzi, H. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 13484–13508.
63. Peng, B.; Li, C.; He, P.; Galley, M.; Gao, J. Instruction Tuning with GPT-4. *arXiv* **2023**, arXiv:2304.03277.
64. Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Ishaan Gulrajani, C.G.; Liang, P.; Hashimoto, T.B. Alpacaeval: An Automatic Evaluator of Instruction-Following Models. 2023. Available online: https://github.com/tatsu-lab/alpaca_eval (accessed on 28 March 2024).
65. Ng, A. The Batch Issue 242: Four Design Patterns for AI Agentic Workflows Blog Post. The Batch. Available online: https://www.deeplearning.ai/the-batch/issue-242/ (accessed on 28 March 2024).
66. Huang, X.; Liu, W.; Chen, X.; Wang, X.; Wang, H.; Lian, D.; Wang, Y.; Tang, R.; Chen, E. Understanding the planning of LLM agents: A survey. *arXiv* **2024**, arXiv:2402.02716.
67. Wei, J.; Yang, C.; Song, X.; Lu, Y.; Hu, N.; Tran, D.; Peng, D.; Liu, R.; Huang, D.; Du, C.; et al. Long-form factuality in large language models. *arXiv* **2024**, arXiv:2403.18802.
68. Church, K.W.; Yue, R. Emerging trends: Smooth-talking machines. *Nat. Lang. Eng.* **2023**, *29*, 1402–1410. [CrossRef]
69. Sierocka, H. Cultural Dimensions Of Legal Discourse. *Stud. Log.* **2014**, *38*, 189–196. [CrossRef]
70. Schilling, T. Beyond Multilingualism: On Different Approaches to the Handling of Diverging Language Versions of a Community Law. *Eur. Law J.* **2010**, *16*, 47–66. [CrossRef]
71. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783.
72. Boginskaya, O. Semantics of the verb shall in legal discourse. *Jezikoslovlje* **2017**, *18*, 305–317.
73. Basmov, V.; Goldberg, Y.; Tsarfaty, R. LLMs' Reading Comprehension Is Affected by Parametric Knowledge and Struggles with Hypothetical Statements. *arXiv* **2024**, arXiv:2404.06283.
74. Zhong, H.; Wang, Y.; Tu, C.; Zhang, T.; Liu, Z.; Sun, M. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 1250–1257. [CrossRef]

75. Zhang, D.; Finckenberg-Broman, P.; Hoang, T.; Pan, S.; Xing, Z.; Staples, M.; Xu, X. Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions. *arXiv* **2024**, arXiv:2307.03941. [CrossRef]
76. Ali, A.; Al-rimy, B.A.S.; Alsubaei, F.S.; Almazroi, A.A.; Almazroi, A.A. HealthLock: Blockchain-Based Privacy Preservation Using Homomorphic Encryption in Internet of Things Healthcare Applications. *Sensors* **2023**, *23*, 6762. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.