Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges

FARID ARIAI, School of Electrical Engineering and Computer Science, The University of Queensland, Saint Lucia, Australia
JOEL MACKENZIE, School of Electrical Engineering and Computer Science, The University of Queensland, Saint Lucia, Australia
GIANLUCA DEMARTINI, School of Electrical Engineering and Computer Science, The University of Queensland, Saint Lucia, Australia

Natural Language Processing (NLP) is revolutionising the way both professionals and laypersons operate in the legal field. The considerable potential for NLP in the legal sector, especially in developing computational assistance tools for various legal processes, has captured the interest of researchers for years. This survey follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses framework, reviewing 154 studies, with a final selection of 131 after manual filtering. It explores foundational concepts related to NLP in the legal domain, illustrating the unique aspects and challenges of processing legal texts, such as extensive document lengths, complex language, and limited open legal datasets. We provide an overview of NLP tasks specific to legal text, such as Document Summarisation, Named Entity Recognition, Question Answering, Argument Mining, Text Classification, and Judgement Prediction. Furthermore, we analyse both developed legal-oriented language models, and approaches for adapting general-purpose language models to the legal domain. Additionally, we identify *sixteen* open research challenges, including the detection and mitigation of bias in artificial intelligence applications, the need for more robust and interpretable models, and improving explainability to handle the complexities of legal language and reasoning.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Applied computing** → **Law**; • **Computing methodologies** → **Natural language processing**; **Artificial intelligence**.

Additional Key Words and Phrases: Natural Language Processing, Artificial Intelligence, Legal Domain, Law

1 Introduction

Advancements in Natural Language Processing (NLP) have significantly impacted the legal domain by simplifying complex tasks, such as Legal Document Summarisation (LDS), Legal Argument Mining (LAM), enhancing legal text comprehension for laypersons, and improving Legal Question Answering (LQA) and Legal Judgement Prediction (LJP) [26, 48, 58, 61, 63, 74, 107, 116]. These improvements – like in many other data-driven fields – are primarily attributed to advancements in Neural Network (NN) architectures, such as transformer models [134]. NLP techniques now enable machines to generate text, answer legal questions, draft regulations, and simulate legal reasoning, which have the potential to revolutionise legal practices [61]. Applications, such as contract review [53, 86, 87, 133] and case prediction [96, 136] have been automated to a large extent, speeding up processes, reducing human error, and cutting operational costs [154]. Additionally, the use of NLP allows lawyers and legal professionals to reduce their workload, enhance their efficiency, and minimise errors in

Authors' Contact Information: Farid Ariai, School of Electrical Engineering and Computer Science, The University of Queensland, Saint Lucia, Queensland, Australia; e-mail: f.ariai@uq.edu.au; Joel Mackenzie, School of Electrical Engineering and Computer Science, The University of Queensland, Saint Lucia, Queensland, Australia; e-mail: joel.mackenzie@uq.edu.au; Gianluca Demartini, School of Electrical Engineering and Computer Science, The University of Queensland, Saint Lucia, Queensland, Australia; e-mail: demartini@acm.org.

ACM 1557-7341/2025/11-ART
https://doi.org/10.1145/3777009

decision-making processes [58, 116]. Despite rapid developments in NLP technology, significant challenges remain in the legal context due to lengthy documents, complex legal language, and complicated document structures [12, 45, 58, 63, 95, 125, 136, 147].

Despite these advantages, the integration of NLP in the legal domain is not without challenges, especially in terms of fairness, bias, and explainability issues [31, 121, 129]. The use of Artificial Intelligence (AI) in legal applications must follow strict standards of accuracy, fairness, and transparency, given the potential impact on clients' lives and rights. Nonetheless, Large Language Models (LLMs) have demonstrated potential to enhance the efficiency, fairness, and precision of legal tasks [39, 58].

This survey article explores the current landscape of NLP applications within the legal domain. It discusses its potential benefits and the practical challenges it poses. NLP is a broad field covering a wide range of techniques for processing, analysing, and understanding human language. By examining the latest advancements and applications of NLP in law, this article provides a comprehensive overview of the field. Table 1 summarises the scope of the survey and categorise the research into several areas: LQA, LJP, Legal Text Classification (LTC), LDS, legal Named Entity Recognition (NER), LAM, legal corpora and legal Language Models (LMs). Each category lists relevant projects and papers, and shows the work being done in each sub-field. Notably, there is comparatively less research in NER and legal corpora, whereas LDS and LQA have seen extensive research activity, with a substantial number of datasets and research contributions. This summary provides an overview of how NLP techniques are applied to various challenges in the legal domain and offers insights into future directions of AI in legal practice.

This document is organised as follows. Firstly, In Section 2, we discuss previous surveys in this multidisciplinary domain. In Section 3, we provide a detailed overview of legal language and the basic principles of NLP as they apply to the legal domain. In Section 4, we briefly explain the research methodology of this work and how we extracted the resources. In sections 5-11, we explore various NLP tasks that are tailored for legal applications, focusing on their unique requirements and the methodologies employed to address them. Additionally, we study the datasets available for training and evaluating legal NLP tasks, emphasising their characteristics and the implications they have for model performance. Following this, in Section 12, we investigate the development of LMs that have been specifically adapted to the legal field. Finally, in Section 13, we address the key challenges associated with deploying NLP technologies in legal settings, discussing both current issues and potential solutions. Since this survey contains many acronyms, Table 2 provides the list of acronyms and their meanings to make it easier to follow.

2 Related Work

Several studies have examined the use of NLP in the legal domain, each focusing on different problems and applications. To provide a comprehensive understanding of the existing research on integrating AI within the legal domain, we present an overview of recent literature reviews, as summarised in Table 3. Most survey papers on intelligent legal systems focus either on traditional NLP technologies for specific tasks, such as LJP and LDS, or adopt a broader approach but still overlook certain applications. As illustrated in Table 3, there is yet to be a comprehensive survey that thoroughly examine all facets of this multidisciplinary field; Our current work aims to bridge this gap by offering a comprehensive survey of all NLP tasks, existing datasets and corpora and LMs in the legal domain.

Dias et al. [33] discussed AI and NLP concepts and their applications in the legal domain. Their study did not analyse legal datasets, specific NLP tasks, or legal LLMs. Our work, in contrast, reviews NLP tasks in the legal domain, including LQA, LDS and LTC, along with legal datasets and corpora. Sun [128] examined two NLP tasks in the legal domain, such as LJP and statutory reasoning. Their study reviewed three datasets and two LLMs related to these tasks. Unlike Sun's work, our survey covers a wider range of NLP tasks and includes

Table 1. An overview of the research areas in legal NLP and the key publications discussed in this survey.

Legal Natural Language Processing		
Language Models	Methods	Li et al. [74], Mamakas et al. [85]
	Models	Al-qurishi et al. [2], Chalkidis et al. [20], Colombo et al. [27], Shi et al. [122], Xiao et al. [140]
Datasets	For Pre-training	Henderson et al. [56], Niklaus et al. [98]
	Benchmarks	Barale et al. [9], Chalkidis et al. [21, 22], Goebel et al. [51], Niklaus et al. [97], Park and James [104], Rabelo et al. [111], Xiao et al. [141], Zheng et al. [151], Östling et al. [157]
Tasks	Argument Mining	Habernal et al. [55], Palau and Moens [101], Poudyal et al. [108], Santin et al. [115]
	Named Entity Recognition	Au et al. [7], Dozier et al. [35], Kalamkar et al. [65], Leitner et al. [72], Păis et al. [110], Smădu et al. [123]
	Document Summarisation	Farzindar and Lapalme [40], Gelbart and Smith [46, 47], Jain et al. [60], Liu et al. [78], Moens et al. [91], Moro et al. [93], Polsley et al. [106], Schraagen et al. [117], Shen et al. [120], Zhong and Litman [156]
	Text Classification	Bambroo and Awasthi [8], Bhattacharya et al. [13], Chalkidis et al. [18, 19], Elnaggar et al. [37], Galassi et al. [43], Grabmair et al. [52], Graham et al. [53], Lee and Lee [71], Mamooler et al. [86], Nguyen et al. [94], Papaloukas et al. [102], Song et al. [124], Tuggener et al. [133], Wang et al. [137]
	Judgement Prediction	Chalkidis et al. [17], Feng et al. [41], Liu et al. [81], Luo et al. [83], Ma et al. [84], Medvedeva et al. [90], Niklaus et al. [96], Semo et al. [118], Tong et al. [132], Xu et al. [142], Yang et al. [143], Ye et al. [144], Zhang et al. [149], Zhong et al. [152, 153]
	Question Answering	Askari et al. [5, 6], Büttner and Habernal [15], Chen et al. [23], Huang et al. [59], Khazaeli et al. [68], Louis et al. [82], Sovrano et al. [125, 126], Yuan et al. [146], Zhang et al. [150], Zhong et al. [155]

Table 2. List of acronyms used in the survey.

Acronyms	Meaning
CFR	Code of Federal Regulations
CJEU	Court of Justice of the European Union
ECHR	European Court of Human Rights
FSCS	Federal Supreme Court of Switzerland
JEC-QA	Judicial Examination of Chinese Question Answering
LAM	Legal Argument Mining
LDS	Legal Document Summarisation
LJP	Legal Judgement Prediction
LQA	Legal Question Answering
LTC	Legal Text Classification
ML-LJP	Multi-Law aware Legal Judgement Prediction

Table 3. An overview of existing surveys on NLP in the legal domain. We use a check mark (✓) to indicate papers that study the most of the existing research on each subject in legal NLP. Papers that do not address a subject receive a cross (✗), and those that partially cover specific subjects are marked with a dash (–).

References	Covered Subjects in the Legal Domain				Published Year
	Dataset	NLP Tasks	LM	Large Corpora	
Dias et al. [33]	–	–	–	✗	2022
Sun [128]	✓	–	✓	✗	2023
Cui et al. [28]	✓	–	✓	✗	2023
Anh et al. [4]	✗	✓	✓	✗	2023
Ganguly et al. [45]	–	–	✓	✗	2023
Chen et al. [25]	✓	✓	✓	✗	2024
Krasadakis et al. [69]	–	–	–	–	2024

legal corpora and datasets. Cui et al. [28] focused on LJP, reviewing 43 datasets in nine languages. Their study evaluated classification, text generation and regression tasks. It also discussed pre-trained LMs used for LJP. Our work differs by covering multiple NLP tasks, legal corpora and dataset availability.

Anh et al. [4] explored challenges in legal language processing and how LLMs can address them. Their study summarised six NLP tasks and discussed ethical concerns, including bias, privacy and transparency. While these issues are relevant, our survey focuses on existing NLP methods, datasets and legal corpora for pre-training and fine-tuning; we discuss elements of bias, fairness, privacy, interpretability, and explainability in Section 13.

Ganguly et al. [45] surveyed legal text processing challenges, such as NER and sentence boundary detection. Their study reviewed historical developments in AI and law research, as well as recent NLP advancements. It covered LDS and LJP but did not examine LQA, LTC or legal corpora. Our work covers all of these areas.

Chen et al. [25] studied LLMs in finance, healthcare and law. While attempting to provide a broad view of LLM applications in the legal domain, their study's scope resulted in a less detailed review of specific NLP tasks and datasets. Additionally, it did not examine large legal corpora or pre-training methods, aspects that our survey addresses.

Krasadakis et al. [69] focused on challenges and advancements in some NLP tasks, such as NER and Relation Extraction. Unlike our study, which reviews all NLP tasks alongside datasets and legal corpora, their work primarily investigated existing LLMs for the legal domain.

The main difference between our work and previous surveys is that our survey aims to provide a more general view of all aspects of NLP tasks in the legal domain, rather than focusing solely on specific applications. The main contributions of this survey are summarised as follows: (1) This article extends previous surveys by examining a broad spectrum of studies and applications of legal NLP. By discussing datasets and large corpora in 24 languages and exploring popular legal LMs, this survey establishes itself as an important resource in the field of legal NLP. (2) The survey offers an in-depth look at the challenges of integrating NLP with legal applications, with detailed discussions on technical solutions that tackle these issues, thereby enhancing understanding and encouraging further research in this evolving field. (3) This survey also highlights the existing research gaps in legal NLP, identifying areas that require further exploration and development and providing a road-map for future research efforts in the legal NLP domain.

3 Background and Foundational Concepts

In this section, we explain how NLP can be applied within the legal domain. We begin by outlining the unique characteristics of legal documents and legal language, highlighting the challenges these features pose for NLP applications. We then introduce core NLP concepts, methods, and paradigms relevant to legal texts, before discussing the recent impact of LLMs in the legal sector. Finally, we summarise key journals, conferences, and workshops that shape the field of legal NLP.

3.1 Legal Documents

Legal documents, such as court filings, judicial judgements, statutes, treaties, contracts, and formal legal correspondence *encode* the authoritative rules, rights, and duties that underpin our legal systems. They transmit legally operative information, including procedural requirements, enforceable obligations, and interpretative reasoning. Lawyers, judges, regulators, and academics consult these sources to analyse cases, interpret legislation, draft or negotiate agreements, verify compliance, and support teaching. Access is typically provided through official court portals, statutory databases, and commercial research platforms.

3.1.1 Legal language and its characteristics. Legal language is characterised by unique features that distinguish it from everyday language, primarily because of its role within the legal system. One prominent feature is its formality, where legal texts often employ a more formal vocabulary and syntax to ensure precision and avoid ambiguity [50]. This formality is important, as the meaning of terms can have legal effects. Specialised vocabularies, fixed syntactic patterns, and even punctuation are chosen to maximise precision and avoid ambiguity, as small drafting choices can decisively alter legal consequences. One such example appeared in the 2006 dispute between Rogers Communications and Bell Aliant,[1] where a single comma in the English version of a termination clause was interpreted to permit early cancellation of a multimillion-dollar pole-access contract, whereas the absence of that comma in the equally-authentic French text preserved the original five-year lock-in. This case illustrates how punctuation alone can shift rights and liabilities. Legal documents also typically use passive constructions and complex sentence structures to provide detailed and comprehensive descriptions [50] without directly attributing actions or intentions to specific parties.

Another distinctive aspect of legal language is its reliance on specialised words and phrases. These include terms with specific legal meanings, archaic words rarely used in everyday language, and standardised phrases embedded in legal tradition [50]. Such language can make legal documents less accessible to non-specialists, necessitating accurate interpretation by legal professionals.

Furthermore, legal language is heavily intertextual, frequently referencing other legal texts, such as statutes, regulations, and case law. This ensures that legal arguments are grounded in existing legal frameworks and previous cases. The extensive use of citations and references situates each document within a wider legal discourse. Such intertextuality requires legal professionals to understand both the texts themselves and the broader legal context in which they operate. To illustrate the intertextuality of legal language, Figure 1 shows a sample page from the Code of Federal Regulations (CFR) of the US, extracted from the Electronic CFR[2], which displays § 40.51, Labour Certification, of 22 CFR. This section is part of Title 22 of the CFR, which governs foreign relations and specifically details the requirements and procedures for labour certification. The text includes underlined references to other legal sources, such as INA 212(a)(5). This citation refers to section 212 of the Immigration and Nationality Act, subsection (a), paragraph (5), which outlines conditions for inadmissibility to the US. In the "Source" section, the citation "56 FR 30422, July 2, 1991," indicates a Federal Register publication: "56 FR" is the volume number, "30422" the page where the document begins and "July 2, 1991" the publication date.

[1] https://crtc.gc.ca/eng/archive/2007/dt2007-75.htm
[2] https://www.ecfr.gov

Editorial Note: Nomenclature changes to part 40 appear at 71 FR 34520 and 34521, June 15, 2006.

§ 40.51 Labor certification.

a. **INA 212(a)(5) applicable only to certain immigrant aliens.** INA 212(a)(5)(A) applies only to immigrant aliens described in INA 203(b)(2) or (3) who are seeking to enter the United States for the purpose of engaging in gainful employment.

b. **Determination of need for alien's labour skills.** An alien within one of the classes to which INA 212(a)(5) applies as described in §40.51(a) who seeks to enter the United States for the purpose of engaging in gainful employment, shall be ineligible under INA 212(a)(5)(A) to receive a visa unless the Secretary of Labor has certified to the Secretary of Homeland Security and the Secretary of State, that

1. There are not sufficient workers in the United States who are able, willing, qualified, (or equally qualified in the case of aliens who are members of the teaching profession or who have exceptional ability in the sciences or the arts) and available at the time of application for a visa and at the place to which the alien is destined to perform such skilled or unskilled labour, and

2. The employment of such alien will not adversely affect the wages and working conditions of the workers in the United States similarly employed.

c. **Labor certification not required in certain cases.** A spouse or child accompanying or following to join an alien spouse or parent who is a beneficiary of a petition approved pursuant to INA 203(b)(2) or (3) is not considered to be within the purview of INA 212(a)(5).

[56 FR 30422, July 2, 1991, as amended at 61 FR 1835, Jan. 24, 1996]

Authority: 8 U.S.C. 1104, 1182, 1183a, 1641

Source: 56 FR 30422, July 2, 1991, unless otherwise noted.

Fig. 1. A sample page from the CFR, illustrating the structured and referenced nature of legal documents.

Disambiguation titles and nested entities are other issues in legal contexts [69]. Disambiguation titles, such as "The President of USA" require precise identification based on contextual details, such as time and location. Nested entities, where titles of legislative articles refer to other laws, introduce further complexity. To further complicate matters, legal documents are frequently provided in non-machine-readable PDF formats, complicating data extraction and processing. Additionally, the variation in legal reasoning, which includes rule-based, analogical and evidentiary arguments, along with changes in legal standards, creates challenges for applying conventional NLP models [42]. Elements such as the peculiar use of punctuation affecting text segmentation and the frequent presence of digits and numbers, can also disrupt traditional NLP pipelines. These challenges demonstrate the need for specialised NLP solutions tailored to the legal domain.

3.1.2 Domains with Shared Characteristics. Other domains exhibit features similar to those of legal texts, such as specialised vocabularies, large-scale corpora and cross-referencing. In the medical domain, texts – including clinical notes, patient records, and research articles – rely on domain-specific terminology, diagnostic labels, and biological taxonomies, often involving case-based analyses akin to legal reasoning. Similarly, software documentation and source code contain specialised programming terms and algorithmic descriptions and are characterised by large repositories with numerous cross-references to libraries and functions.

3.2 Legal NLP

The legal sector has been exploring AI-driven solutions since the late 20th century, applying NLP techniques to automate legal processes. NLP applications in the legal field include drafting client briefs and analysing large document sets, enabling smaller firms to compete more effectively with larger ones [89]. These applications are also very important for compliance and due-diligence checks – required when companies merge their business,

for instance – and greatly supports legal education and learning in fast-changing fields [89]. Legal NLP can also enhance the analysis of complex legal documents, thereby assisting with complex decision-making processes [58].

The foundation of NLP is text and the legal domain primarily consists of textual data [3], including statutes, case law, contracts and regulations. Given the text-intensive nature of the legal field, NLP offers potential to change how legal professionals interact with and use this information. By leveraging advanced algorithms and Machine Learning (ML) models, legal NLP aims to make legal texts more accessible, interpretable and actionable [58].

3.2.1 Basic foundations and concepts of NLP. The integration of NLP in the legal domain relies on foundational techniques that enable the processing and analysis of legal texts. These techniques form the basis for various applications, transforming unstructured legal documents into structured, actionable information. This section introduces key NLP methods, ML paradigms and text retrieval technique.

(1) Core NLP methods:
- `Tokenization`: Tokenization is the process of breaking text into smaller units, typically words or subwords, known as tokens. It is a fundamental step in NLP, allowing structured or unstructured text to be analysed. In legal NLP, tokenization facilitates processing lengthy documents by segmenting them into manageable parts and prepares the input text for numerical representation through word embeddings.
- `Word Embeddings`: Embeddings represent words as continuous (numeric) vectors in a high-dimensional space, designed to capture semantic relationships. These embeddings enable models to encode word meanings and relationships, which are essential for tasks such as legal text similarity analysis and document classification.
- `Transformers`: The transformer architecture is a NN model designed to process text by first identifying connections between input words using self-attention. In this process, it computes attention weights (which determine the importance of each word based on its context) across the entire input sequence before generating output representations. Transformers operate on word embeddings and refine these embeddings by incorporating context from the entire sequence.
- `PLMs`: PLMs, such as Bidirectional Encoder Representations from Transformers (BERT) [32], are transformer-based models pre-trained on large text corpora using self-supervised objectives. These models learn linguistic patterns and can be fine-tuned for specific legal NLP tasks. PLMs leverage transformer architectures and pre-trained embeddings to perform NLP tasks.

(2) ML paradigms for NLP:
- `Multi-task Learning (MTL)`: MTL is an approach where a model learns multiple related tasks simultaneously, leveraging shared knowledge across tasks. This technique improves model robustness and efficiency, particularly in data-scarce legal NLP applications [24].
- `Parameter-Efficient Fine-Tuning (PEFT)`: PEFT is a method for adapting PLMs to new tasks that involves freezing the majority of the model's parameters, relying on updating just a small subset "fine-tuned" to the downstream task. This approach significantly reduces the computational resources and time required for fine-tuning, making it particularly effective in resource-limited scenarios, while still achieving competitive performance in tasks, such as text generation [75].

These ML paradigms enhance the performance of core NLP methods, particularly PLMs, by tailoring models to specific tasks and optimising their training efficiency.

(3) Text retrieval technique:
- `Retrieval-Augmented Generation (RAG)`: RAG combines traditional Information Retrieval (IR) methods with generative NLP models, allowing systems to retrieve external knowledge before generating responses.

Text retrieval techniques, such as RAG complement the core NLP methods by providing additional context and external information that enhances the generation and refinement of texts.

3.2.2 LLMs as an application of NLP in the legal sector. LLMs are a category of Deep Learning (DL)-based NLP models trained on extensive text corpora to process and generate human-like language. These models, typically based on Transformer architectures, learn linguistic patterns from diverse sources, enabling them to perform tasks, such as text summarisation, document analysis and Question Answering (QA). LLMs have gained widespread attention in NLP applications, particularly following the release of ChatGPT in November 2022 [99].

A notable demonstration of NLP's potential in the legal sector occurred when GPT-4 was reported to have passed the Uniform Bar Exam near the 90th percentile, underscoring the technology's potential [67]. However, subsequent analyses by Martínez [88] suggest that its actual performance may be considerably lower, possibly around the 48th percentile overall and 15th percentile on essays. Similarly, further research reveals that although ChatGPT can achieve moderate success on certain legal classification tasks, smaller fine-tuned models still outperform it by about 30 percentage points [16]. Another study highlights a tendency for LLMs to hallucinate legal content at high rates—up to 58% in some cases, raising reliability concerns for complex tasks [29]. Despite these drawbacks, lawyers and law students remain conscious of the broader impact of such models: a recent LexisNexis survey [73] shows that about half of all lawyers believe LLMs will transform legal practice, with 92% anticipating at least some impact. Furthermore, 77% of respondents foresee efficiency gains for legal professionals and 63% predict changes in how law is taught and studied.

3.2.3 Key publications and conferences in legal NLP. This section highlights the key journals, conferences and workshops that serve as platforms for sharing advancements and insights at the intersection of NLP and the legal domain. These resources provide opportunities for researchers to engage with cutting-edge work in legal NLP.

Several leading journals focus on the intersection of AI, NLP and the legal domain. *"Artificial Intelligence and Law,"* published by Springer, is a leading journal that features research articles on legal reasoning, legal IR and legal knowledge representation. A recent special issue, Applications and Evaluation of LLMs in the Legal Domain, examines the use of LLMs in legal tasks such as summarisation, judgement prediction and contract drafting, while also addressing concerns, such as bias, misinformation and regulatory compliance. Additionally, the *"Journal of Law and Information Technology"* focuses on the application of information technology in law, including research AI.

Conferences significantly advance research and promote collaboration in legal NLP. The International Conference on Artificial Intelligence and Law (ICAIL) is a biennial event showcasing advances in AI applications for the legal domain, including NLP and ML. The *Conference on Legal Knowledge and Information Systems* (JURIX) is an annual event that focuses on legal informatics and NLP technologies. In addition to these dedicated venues, prominent legal NLP research has also been published in broader AI, NLP and IR conferences, including NeurIPS, ACL (and its associated workshops), NAACL, EMNLP, SIGIR, IJCAI, AAAI and LREC.

In legal NLP, workshops also attract strong research contributions. The workshop on *Automated Semantic Analysis of Information in Legal Texts* focuses on NLP and semantic analysis of legal documents. The *International workshop on Juris-Informatics* (JURISIN) brings together researchers from law, social science and technology to discuss foundational and practical issues at the intersection of legal theory and informatics. The *Natural Legal Language Processing* (NLLP) workshop provides a platform for discussing NLP technologies tailored for legal texts and is often part of major NLP conferences. The *EXplainable AI in Law* (XAILA) workshop focuses on the explainability of AI systems in legal contexts, aiming to improve transparency and trust in AI applications. The *Competition on Legal Information Extraction/Entailment* (COLIEE) is an annual event that challenges participants to develop innovative solutions for legal information extraction and entailment tasks. Additionally, the *Legal Track* at the Text Retrieval Conference (TREC), which ran from 2006 to 2011, focused on evaluating IR techniques for legal document review, providing a venue for researchers to test and compare methods in the context of

e-discovery. Its contributions include benchmark datasets and evaluation metrics that were used to assess retrieval performance in legal text processing.

4 Methodology

This survey follows the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) framework [100]. It ensures a transparent and comprehensive assessment of research on NLP tasks within the legal sector.

4.1 Search Strategy

We performed a systematic search across two academic databases to identify relevant studies, including: Google Scholar and IEEE Xplore. Then, search queries were crafted to capture studies that focused on the application of NLP to legal tasks. The search was defined by the following two queries:

- Query 1: ("Natural Language Processing" OR "NLP") AND ("Legal" OR "Law")
- Query 2: ("Legal" AND ("Named Entity Recognition" OR "NER" OR "Document Summarisation" OR "Text Classification" OR "Document Classification" OR "Judgement Prediction" OR "Question Answering" OR "Corpus" OR "Language Model" OR "Argument Mining"))

Our search covered publications within the following date ranges for each NLP task: LQA from 2020-2024, LJP from 2017-2024, LTC from 2018-2023, LDS from 2016-2024, legal NER from 2010-2022, and LAM from 2009-2024. Furthermore, we limited our search for legal corpora to 2021-2024 and legal LMs from 2020-2024. This approach ensured the inclusion of recent advancements. Peer-reviewed journal articles and high-quality conference proceedings were prioritised, with secondary consideration given to relevant non-peer-reviewed sources.

4.2 Study Selection

A total of 154 studies were initially identified from the database search. To refine this list, we applied manual review. This process involved:

(1) **Title and Abstract Screening:** We reviewed the titles and abstracts of all retrieved studies to assess their relevance to the predefined legal NLP tasks. Studies unrelated to the core legal NLP and its tasks were excluded.
(2) **Full-Text Review:** Articles that passed the initial screening underwent a detailed full-text review to confirm their relevance, quality and alignment with the inclusion criteria. During this phase, we also examined the literature review sections of each paper to ensure that the studies not only contributed original findings but also demonstrated a comprehensive understanding of the existing legal NLP research landscape.
(3) **Final Selection:** Of the original 154 studies, 131 met the inclusion criteria and were retained. These studies were selected for their direct relevance to key legal NLP tasks, methodological quality, and engagement with existing literature.

4.3 Eligibility Criteria

To determine which studies were included in the final synthesis, we established the following criteria:

- **Inclusion Criteria**:
  - The study must focus on at least one of the target NLP tasks (LQA, legal NER, LJP, LDS, LTC, LAM), or it must focus on legal LMs or legal corpora.
  - The study must present empirical research or significant methodological contributions to legal NLP.
  - Both peer-reviewed and non-peer-reviewed studies were considered if they provided valuable insights.

- **Exclusion Criteria**:
  - Studies focused exclusively on unrelated areas such as IR methods, pattern mining, information extraction or similarity detection without a clear application to the specific legal NLP tasks mentioned.
  - General NLP studies without a focus on legal applications.
  - Editorials, opinion pieces or other non-research articles.
  - Papers that did not meet basic methodological standards were not included in the final analysis.

5 Legal Question Answering

LQA involves responding to legal queries, a task typically performed by professionals with domain expertise. It requires a comprehensive review of relevant laws, careful interpretation of statutes and regulations, and the application of legal principles and precedent to the facts. LQA aims to provide legal advice, helping individuals and businesses navigate the complex legal landscape.

5.1 Datasets

LQA datasets are a specialised resource designed to facilitate research in the domain of legal NLP. They consist of a collection of legal questions and corresponding answers, drawn from various legal documents and case law. Most questions in the LQA datasets fall into two main categories: knowledge-driven questions (KD-questions) and case-analysis questions (CA-questions) [155]. KD-questions are centred around the understanding of specific legal concepts, whereas CA-questions involve the analysis of actual legal cases. Both categories demand advanced reasoning skills and a deep comprehension of the text, making LQA a particularly challenging task in the field of NLP.

Zhong et al. [155] introduce JEC-QA, a dataset with 26,365 multiple-choice questions from the National Judicial Examination of China and related websites. Each question provides four possible answers and is annotated with the type of reasoning required, such as word matching, conceptual understanding, numerical analysis, multi-paragraph comprehension and multi-hop inference. This dataset poses challenges for QA models, highlighting the gap between machine performance and human expertise in legal reasoning.

Sovrano et al. [125] present the Q4PIL dataset, designed to evaluate automated QA systems in the domain of Private International Law. It includes 17 carefully selected questions based on key EU regulations – Rome I, Rome II and Brussels I bis – with answers derived directly from these regulations. The questions are classified based on their specificity, allowing for nuanced analysis of context-dependency in legal reasoning. This dataset supports the assessment of QA systems intended for legal professionals navigating complex cross-border issues.

EQUALS [23] is a large-scale annotated LQA dataset in Chinese law, containing 6,914 question-answer pairs with answers based on specific law articles. Curated by senior law students, it covers 10 collections of Chinese laws and includes annotations indicating the type of reasoning required for each question. The dataset ensures that answers are precise excerpts from relevant law articles, making it valuable for developing advanced LQA systems that can aid in legal research and decision-making.

Büttner and Habernal [15] introduce GerLayQA, a dataset supporting LQA for laypersons in Germany, focusing on the civil-law system. It contains 21,538 real-world questions posed by laypersons, paired with expert answers from lawyers grounded in specific paragraphs of German legal codes. The dataset was constructed through filtering and quality assurance to ensure accuracy and relevance, making it a valuable resource for developing LQA systems that interpret and apply German law to everyday legal inquiries.

5.2 Approaches

Recently, DL has been applied in LQA through NN models trained on large datasets to identify complex patterns and relationships. These models analyse posed questions, recognise relevant legal topics, and generate appropriate answers based on learned patterns and memory.

Modern ML approaches to LQA use NN architectures to process natural language. Popular architectures include Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN), which can be fine-tuned for QA tasks. These models adapt to new patterns, capture contextual information and generate more accurate responses. Transformer-based models, such as BERT and, more recently, ChatGPT, have proven particularly effective in NLP tasks. These models use the transformer architecture and self-attention mechanisms to learn the context of text to model the patterns and word dependencies in text. This allows them to provide relevant answers by weighting the importance of different parts of the input based on its contextual importance. In the following paragraphs, we will study the existing LQA works in the legal domain.

Huang et al. [59] introduce the Artificial Intelligence Law Assistant, the first Chinese LQA system that integrates a legal Knowledge Graph (KG) to enhance query comprehension and answer ranking. The system collects a large-scale QA corpus from an online legal forum and constructs a legal KG with over 42,000 legal concepts. It employs a knowledge-enhanced interactive attention network using Bidirectional LSTM (Bi-LSTM) and co-attention mechanisms to enrich semantic representations of question–answer pairs with legal domain knowledge. Additionally, it provides visual explanations for selected answers, offering users a clear understanding of the QA process.

Khazaeli et al. [68] develop an IR-based QA system tailored to the legal domain, combining sparse term-based search (BM25) and dense vector techniques (semantic embeddings) as input to a BERT-based answer re-ranking module. The system utilises Legal GloVe and Legal Siamese BERT embeddings to enhance retrieval performance. An "answer finder" component computes the probability that a passage answers the question using a BERT sequence classifier fine-tuned on question–answer pairs, thereby enhancing the model's ability to discriminate relevant answers.

Li et al. [76] introduce a retrieve-then-answer framework featuring a Graph-Based Evidence Retrieval and Aggregation Network (GESAN) to enhance LQA on the JEC-QA dataset [155]. The framework leverages legal knowledge by predicting question topics and retrieving relevant paragraphs using BM25. GESAN aggregates the evidence and processes it along with the question and options to generate accurate predictions, demonstrating improved reasoning capabilities in LQA.

Askari et al. [5] tackle legal expert finding on QA platforms by building query-dependent textual profiles for lawyers. Using data from the Avvo forum,[3] they represent each lawyer with four facets – comment content, positive sentiment, negative sentiment, and answer recency – derived from past answers and comments. A separate BERT model is fine-tuned for each facet and the scores are linearly combined with a document-based BERT ranker to produce the final ranking.

Zhang et al. [150] re-frame LQA as a generation task with GLQA, a retrieve-then-generate framework that runs both retrieval and generation in a single T5 model via MTL. A knowledge retriever encodes questions and law articles into dense embeddings, then retrieves the top-k relevant statutes. These articles are concatenated with the question and passed to a knowledge-enhanced generator that produces an answer grounded in the retrieved text.

Louis et al. [82] propose an end-to-end "retrieve-then-read" system that generates long-form answers to statutory questions. A lightweight bi-encoder first retrieves the most relevant French legal provisions, after which an instruction-tuned LLM, adapted via in-context learning and PEFT, composes detailed answers that cite those

[3]https://www.avvo.com

statutes. To ensure transparency, the model also outputs extractive rationales, listing the exact paragraphs that justify each response.

Sovrano et al. [126] propose DiscoLQA, a discourse-based LQA system that focuses on important discourse elements, such as Elementary Discourse Units and Abstract Meaning Representations. This approach helps the answer retriever identify the most relevant parts of the discourse, enhancing retrieval accuracy. They introduce the Q4EU dataset, containing over 70 questions and 200 answers on six European norms, demonstrating improved performance in LQA even without domain-specific training.

Yuan et al. [146] present a three-step approach to bridge the legal knowledge gap by creating CLIC-pages—snippets that explain technical legal concepts in layperson's terms. They construct a legal question bank containing legal questions answered by CLIC-pages, using GPT-3 [14] to generate machine-generated questions. The study shows that machine-generated questions improve scalability and access to legal information for non-experts.

Askari et al. [6] propose a cross-encoder re-ranker (CE_{FS}) for legal answer retrieval, incorporating fine-grained structured inputs from community QA data to enhance retrieval performance. They introduce the LegalQA dataset containing 9,846 questions and 33,670 lawyer-curated answers. The approach involves a two-stage ranking pipeline with a BM25 retriever followed by a re-ranker, showing that integrating question tags into the input structure can bridge the knowledge gap and improve retrieval in the legal domain.

6 Legal Judgement Prediction

LJP is a key task in legal NLP, particularly in civil-law jurisdictions where judgements rely on case facts and statutory provisions [154]. It seeks to predict legal outcomes from case descriptions and the applicable legislation [154]. The topic has attracted growing interest from AI researchers and legal professionals because of its potential to help judges, lawyers and scholars anticipate case results based on historical data [25].

LJP is clearly a demanding and complex problem. Historical legal data can contain inherent biases that can create feedback loops and amplify discrimination if not managed carefully [69]. Therefore, ensuring the impartiality of predicted rulings is crucial [69]. At present, LJP is still performed chiefly by legal experts who require extensive specialised training to identify relevant statutes, define charge ranges, and set penalty terms [28].

6.1 Datasets

The LJP datasets are specialised resources designed to advance research in predicting judicial outcomes within the domain of legal NLP. These datasets are categorised into four main types: court view generation, law articles, charge prediction and prison term prediction. Court view datasets contain judicial opinions and summaries. Law articles datasets focus on predicting outcomes based on specific statutes or regulations. Charge prediction datasets aim to identify the charges appropriate to the case details. Prison term datasets estimate likely sentence durations given the crime and legal context. Each category presents unique challenges, requiring not only text comprehension but also the ability to apply complex legal reasoning.

The Court View Gen [144] dataset is an innovative resource containing 171,981 Chinese legal cases, each involving a single defendant and a corresponding charge, covering a total of 51 charge categories. It is specifically curated to support the generation of court opinions based on charge labels. All cases were collected from the publicly available *China Judgements Online* repository.

Niklaus et al. [96] introduce a multilingual LJP dataset from the Federal Supreme Court of Switzerland (FSCS), containing over 85,000 cases in German, French, and Italian. The dataset is annotated with publication years, legal areas, and cantons of origin, making it suitable for NLP applications in judgement prediction.

Semo et al. [118] introduce the first LJP dataset centred on US class-action lawsuits. Unlike prior work that relies on court-written summaries of facts, this dataset targets outcome prediction directly from plaintiffs' complaints. A rule-based extraction system identifies the relevant text spans within each complaint.

6.2 Approaches

Luo et al. [83] propose an attention-based NN to enhance charge prediction by jointly modelling charge prediction and relevant law article extraction. They used Bidirectional Gated Recurrent Units (Bi-GRUs) to encode fact descriptions and an article extractor to identify top relevant law articles. The model employs an attention mechanism guided by context vectors to combine embeddings for prediction. Evaluations on Chinese judgement documents showed improved accuracy in predicting charges and providing relevant legal articles.

Zhong et al. [152] introduce TopJudge, a topological MTL framework that models dependencies among subtasks in LJP, such as law article prediction, charge prediction and penalty terms. Using a directed acyclic graph, TopJudge processes subtasks in a topological order reflecting real-world legal decision-making. Evaluated on large-scale Chinese criminal case datasets, it outperformed previous models in predicting legal outcomes.

Ye et al. [144] tackle court view generation from criminal case fact descriptions to improve the interpretability of charge prediction systems and support automatic legal document drafting. They frame the task as a text-to-text natural language generation problem, employing a label-conditioned sequence-to-sequence model with attention to generate court views from encoded charge labels. Their approach advances automatic legal document generation by providing explicit justifications for charge decisions.

Yang et al. [143] propose a Multi-Perspective Bi-Feedback Network (MPBFN) with a Word Collocation Attention mechanism to improve LJP. The MPBFN addresses the challenges of multiple subtasks and their dependencies by using a bi-feedback mechanism for forward prediction and backward verification among subtasks. The Word Collocation Attention integrates word collocation features and numerical semantics to better predict penalties. Evaluated on the CAIL datasets [141], their model outperformed baselines in predicting law articles, charges, and penalty terms.

Chalkidis et al. [17] introduce an English LJP dataset containing approximately 11,500 cases from the European Court of Human Rights (ECHR). They evaluated various neural models on this dataset, including a hierarchical version of BERT (HIER-BERT) to handle long legal documents. Their models outperformed previous feature-based approaches in tasks, such as violation classification and case importance prediction. They also explored potential biases in legal predictive models using data anonymization.

Medvedeva et al. [90] use a linear Support Vector Machine (SVM) to predict whether the ECHR will find a violation of any of nine Convention articles. Trained on the textual proceedings, the model achieves an average 75% accuracy, yet performance drops to 58–68% when predicting future cases. They also found that predicting outcomes based solely on judges' surnames could achieve accuracy of 65%, highlighting potential biases in LJP data.

Zhong et al. [153] introduce QAjudge, a reinforcement learning (RL)-based model designed to provide interpretable legal judgements by visualising the prediction process. QAjudge uses a Question Net to iteratively select relevant yes-no questions about case facts, an Answer Net to provide answers and a Predict Net to generate the final judgement. The model aims to minimise the number of questions asked. Evaluated on real-world datasets, QAjudge demonstrated potential in providing reliable and transparent legal judgements.

Xu et al. [142] propose the Law Article Distillation based Attention Network (LADAN), an end-to-end model addressing the issue of confusing charges in LJP by distinguishing similar law articles. The model uses a novel graph NN to learn differences between confusing law articles and an attention mechanism to extract discriminative features from fact descriptions. Experiments on real-world datasets showed that LADAN improved performance over previous methods in law article prediction, charge prediction and penalty term prediction.

Ma et al. [84] introduce MSJudge, a MTL framework designed to predict legal judgements by leveraging multi-stage judicial data, including pre-trial claims and court debates. MSJudge consists of components to encode multi-stage context, model interactions among claims, facts and debates and predict judgements. Evaluated on a

large civil trial dataset, MSJudge more accurately characterises the interactions among claims, facts and debates for judgement prediction, achieving improvements over state-of-the-art (SOTA) baselines.

Feng et al. [41] address limitations of SOTA LJP models by proposing an event-based prediction model with constraints to improve performance. The model extracts fine-grained key events from case facts and predicts judgements based on these events rather than the entire fact statement. They manually annotated a legal event dataset and introduced output constraints to guide learning. Their method leverages event information and cross-task consistency constraints.

Tong et al. [132] introduce GJudge, a graph boosting framework incorporating constraints to address shortcomings of traditional LJP methods. GJudge features a multi-perspective interactive encoder and a Multi-Graph Attention Network (MGAT) consistency expert module. The encoder integrates fact descriptions with label similarity connections, while the expert module differentiates similar labels and preserves task consistency. Experiments on the CAIL datasets show that GJudge outperforms other models, including the SOTA RLJP [139], achieving higher F1 scores.

Previous works mainly focused on creating accurate representations of a case's fact description to enhance judgement prediction performance. However, these methods often overlook the practical judicial process, where human judges compare similar law articles or potential charges before making a decision. To address this gap, Zhang et al. [149] propose CL4LJP, a supervised contrastive learning framework to improve LJP by capturing fine-grained differences between similar law articles and charges. The framework includes contrastive learning tasks at the article, charge and label levels, enhancing the model's ability to model relationships between fact descriptions and labels.

Liu et al. [81] propose ML-LJP, a multi-law aware LJP method that expands law article prediction into a multi-label classification task incorporating both charge-related and term-related articles. The approach uses label-specific representations and contrastive learning to distinguish similar definitions. A Graph Attention Network (GAT) is employed to learn interactions among multiple law articles for prison term prediction. Experiments showed that ML-LJP outperformed SOTA models, particularly in prison term prediction.

7 Legal Text Classification

LTC is an important task within the domain of NLP that involves categorising legal documents based on their content, a foundational aspect of building intelligent legal systems. With the exponential growth of legal documents, it has become increasingly challenging for legal professionals to locate relevant rulings in similar cases for argumentation. LTC addresses this challenge by automatically associating legal texts with predefined categories, such as criminal, civil or administrative cases, thereby simplifying legal research and decision-making processes.

In the legal field, this process is often referred to as predictive coding, where ML algorithms are trained through supervised learning to classify documents into specific categories. The broader task of text classification in NLP involves assigning one or multiple categories to a document from a set of predefined options and it can take various forms, including binary classification (predicting whether a document is in a given class or not), multi-class classification (predicting which one of many classes a document belongs to) and multi-label classification (predicting which set of classes a document belongs to). Legal document classification often falls under large multi-label text classification, where the label space can consist of thousands of potential categories, adding complexity to the task [119].

7.1 Datasets

LTC datasets are characterised by their domain-specific vocabulary and multi-label nature, requiring models to interpret complex legal texts and categorise them into single or multiple legal themes.

Chalkidis et al. [18] release EURLEX57K, a dataset containing 57,000 EU legislative documents from the EUR-Lex portal[4], annotated with EuroVoc[5] concepts. This dataset facilitates research in LTC, including extreme multi-label text classification, few-shot and zero-shot learning, with documents tagged with an expansive set of descriptors.

Tuggener et al. [133] introduce LEDGAR, a multi-label corpus of legal provisions from contracts scraped from the the US Securities and Exchange Commission's website. The dataset includes over 846,000 provisions across 60,540 contracts, with an extensive label set suitable for text classification and legal studies.

Chalkidis et al. [19] present MULTI-EURLEX, a multilingual dataset containing 65,000 EU laws translated into 23 official EU languages, annotated with EuroVoc labels. The dataset emphasises temporal concept drift by adopting chronological splits, enhancing its utility for sophisticated LTC tasks requiring understanding legal terms across different time periods.

Papaloukas et al. [102] introduce the Greek Legal Code dataset, categorising approximately 47,000 Greek legislative documents into a detailed multi-level classification system. The dataset is structured into volumes, chapters and subjects, each containing diverse legal documents from Greek legislation history, supporting LTC in the Greek legal domain.

Song et al. [124] introduce POSTURE50K, a legal dataset containing 50,000 the US legal opinions annotated with Legal Procedural Postures ranging from common to rare motions. The dataset includes an innovative split strategy to support supervised and zero-shot learning evaluations, ensuring infrequent categories are adequately represented, enhancing model generalizability and testing accuracy.

Graham et al. [53] develop a domain-specific dataset for LTC that focuses on deontic modalities in contract sentences. They manually annotated sentences from the CUAD dataset [57] for permissions, obligations and prohibitions, providing a resource for modelling and studying these functional categories in legal analysis.

Galassi et al. [43] extend the Claudette[6] corpus from 25 to 50 ToS per language (English, German, Italian, and Polish). Each sentence is labelled according to nine categories of potential unfairness, with annotations indicating the degree of unfairness. The dataset was carefully compiled based on language availability, structural similarity and version correspondence. Cross-lingual analysis revealed notable differences between languages – particularly in German – including variations in length, structure, missing clauses, and legal terminology that reflect manual drafting adjustments rather than simple automated translations.

7.2 Approaches

DL methods typically requires extensive data to yield effective results, but MTL can help mitigate data scarcity. Elnaggar et al. [37] leverage transfer learning and MTL to perform tasks such as translation and multi-label classification within legal document corpora. They employ theMultiModel algorithm [64], a fully convolutional sequence-to-sequence architecture that integrates multiple modality networks. The model maps legal texts into a shared embedding space, enabling task switching and improving generalisation across tasks to make efficient use of limited legal data.

Lee and Lee [71] examine LTC in Korean by comparing three DL architectures: CNN with ASCII encoding, CNN with Word2Vec embeddings, and RNN with Word2Vec embeddings. Each model assigns case documents to civil, criminal, or administrative categories. Using a dataset of nearly 60,000 past case documents, the study finds that the RNN model with Word2Vec embedding achieves the highest classification accuracy.

Bambroo and Awasthi [8] introduce an architecture that integrates long attention mechanisms with a distilled BERT model pre-trained on legal domain-specific corpora. Their model employs a combination of local windowed attention and task-motivated global attention to handle inputs up to eight times longer than standard BERT models.

[4]https://eur-lex.europa.eu
[5]https://op.europa.eu/en/web/eu-vocabularies
[6]https://claudette.eui.eu

The architecture, based on the lightweight DistilBERT transformer [114], and incorporating LongformerSelf-Attention, is optimised for legal document classification, outperforming a fine-tuned BERT model and other transformer-based models in both speed and performance.

Song et al. [124] present a DL-based system built on top of RoBERTa [80] for multi-label legal document classification. They enhance the model with domain-specific pre-training, a label-attention mechanism and MTL to improve classification accuracy, particularly for low-frequency classes. The label-attention mechanism uses label embeddings to bridge the semantic gap between samples and class labels, addressing class imbalance issues.

Wang et al. [137] introduce a Document-to-Graph Classifier to classify legal documents based on facts and reasons rather than topics. They extract key entities and represented legal documents using four distinct relation graphs capturing different aspects of entity relationships. A graph attention network [135] is used to learn document representations from the combined graph, improving classification by focusing on factual content.

Mamooler et al. [86] propose an active learning pipeline to fine-tune PLMs for LTC, thereby addressing the challenges of a specialised vocabulary and high annotation costs. Their method involves continual pre-training of RoBERTa on legal texts, knowledge distillation using a pre-trained sentence transformer, and an initial sampling strategy based on clustering unlabelled data. This approach reduces the number of labelling actions required for LTC tasks, reducing the overall cost of the LTC process.

Grabmair et al. [52] introduce LUIMA, a system designed for conceptual legal document retrieval, focusing on vaccine injury decisions. LUIMA employs a multi-level classification pipeline: rule-based sub-sentence annotations tag legal concepts, such as terms and mentions, while ML classifiers categorise sentences into argumentative roles, such as legal rules or evidence-based findings. These annotations feed into a sentence-level indexing process using Apache Lucene, enabling semantic querying beyond traditional keyword search. A learning-to-rank module refines the retrieved results by leveraging hand-crafted features such as sentence match counts and term similarity scores.

Galassi et al. [43] explore the binary classification task of detecting potentially unfair clauses in online Terms of Service (ToS) at the sentence level. Sentences are labelled as either unfair, potentially unfair, or fair. Starting from an English-trained model, the authors compare four cross-lingual approaches: (1) training separate models on each language, (2) projecting annotations from English to another language, (3) translating English training documents and using their original annotations and (4) translating query documents into English during prediction. The evaluation shows scenario (4) achieves similar or better results than scenario (1) when translation quality is high. If the translation quality is low, scenario (1) can produce better results. Scenarios (2) and (3) perform slightly below scenario (1) but remain viable alternatives.

Rhetorical role labelling is the task of classifying sentences in legal documents based on their functional roles, such as fact, argument or ruling, to structure and analyse judicial decisions. The Artificial Intelligence for Legal Assistance (AILA) shared task series focuses on advancing legal NLP by introducing datasets and challenges that address core legal text processing tasks. The latest edition, AILA 2021 [103], featured a rhetorical role labelling task that required classifying sentences into seven predefined roles. Building on these challenges, DeepRhole [13] introduces a transformer-based framework that fine-tunes domain-adapted models such as LegalBERT to capture the semantics of legal texts. The system applies different word embedding strategies, including Law2Vec and embeddings from Google News, to model legal language variability. An inter-annotator study examines the subjectivity in rhetorical role assignment, further assessing the model's performance in different judicial contexts.

8 Legal Document Summarisation

LDS is a specialised branch of automatic summarisation that condenses legal texts, such as court judgements, into clear and informative summaries. Unlike general text summarisation, which extracts key details without following specific formatting rules, LDS must accommodate the distinct structure and specialised content of

legal documents. These documents often include complex elements that are essential for presenting the legal arguments and decisions accurately, such as article numbers, statutory language, and citations. The inherent complexity of legal texts, characterised by their length and detailed internal structures (sections, articles, and paragraphs), demands customised summarisation techniques. This requirement is reinforced by the hierarchical significance of documents based on judicial origin, where interpretations may differ between higher and lower court opinions [66].

LDS can be approached through extractive and abstractive methods. Extractive techniques identify and select the most important sentences or phrases directly from the text, preserving original wording and meaning. Abstractive methods, by contrast, generate new sentences that paraphrase the key information, aiming for conciseness while maintaining the essence of the legal text.

8.1 Datasets

LDS datasets are largely built from structured court proceedings and decisions, providing rich sources for both extractive and abstractive summarisation methods. These datasets often use abstractive summarisation to achieve concise, readable summaries that transform the original legal language into more accessible forms [120].

Shen et al. [120] introduce Multi-LexSum, an abstractive summarisation dataset tailored for the US federal civil rights lawsuits, containing 40,000 source documents and 9,000 expert-written summaries of diverse lengths.

Liu et al. [78] published the Common Law Court Judgement Summarisation (CLSum), a dataset designed for summarising multi-jurisdictional common law court judgements from Australia, Hong Kong, the United Kingdom, and Canada. This dataset utilises LLMs for data augmentation and incorporates legal knowledge to enhance summary generation and evaluation. This dataset addresses the challenge of sparse labelled data in legal domains. CLSum includes a collection of judgements and summaries from prominent court websites. They employ a two-stage summarisation process with techniques such as sparse attention mechanisms and efficient training methods to process lengthy legal documents with limited computational resources.

8.2 Approaches

Several systems have been specifically designed to summarise legal documents. One of the first systems in this field was the Fast Legal EXpert CONsultant (FLEXICON), created by Gelbart and Smith in 1991 [46]. FLEXICON utilises a keyword-based approach [47], scanning a database of terms to pinpoint crucial segments of text. Following this, Moens et al. [91] introduced the SALOMON system in 1999, which employs cosine similarity to cluster similar text regions, aiming to highlight relevant topics within the documents. This method aligns with other abstraction-oriented techniques seen in the work of Erkan and Radev [38]. LetSum [40], developed by Farzindar and Lapalme in 2004, also adopts a keyword-centric strategy but uses “cue phrases” to identify text related to specific themes such as ‘Introduction’, ‘Context’ and ‘Conclusion’. Although LetSum approximated human-written summaries reasonably well, it often produced documents that were longer than desired.

Building on previous developments in LDS, Polsley et al. [106] introduce Casesummarizer, a tool designed for the legal domain that pre-processes legal texts into sentences, scores them using a TF-IDF matrix from extensive legal case reports and enhances sentence scoring by identifying entities, dates and section headings. The tool provides a user-friendly interface with scalable summaries, lists of entities and abbreviations and a significance heat map.

Nguyen et al. [94] propose an RL framework to enhance deep summarisation models for the legal domain, utilising Proximal Policy Optimisation with a reward function that integrates both lexical and semantic criteria. They fine-tune an extractive summarisation backbone based on BERTSUM [79], employing a reward model that includes lexical, sentence, and keyword-level semantics to produce better legal summaries. Schraagen et al. [117] apply an RL approach with a Bi-LSTM and a DL approach based on the BART transformer model

to abstractive summarisation of the Dutch case verdict database Rechtspraak.nl, combining extractive and abstractive summarisation to retain core facts while creating concise summaries.

Zhong and Litman [156] focus on extractive summarisation of legal case decisions, proposing an unsupervised graph-based ranking model that leverages a re-weighting algorithm to utilise document structure properties. They extend the HipoRank model [34] with a novel re-weighting algorithm to improve sentence selection, reducing redundancy and enhancing the inclusion of argumentative sentences from underrepresented sections.

Moro et al. [93] introduce a transfer learning approach that combines extractive and abstractive summarisation techniques to address the lack of labelled legal summarisation datasets, outperforming previous results on the Australian Legal Case Reports dataset and establishing a new baseline for abstractive summarisation.

Jain et al. [60] propose a sentence scoring approach, DCESumm, which combines supervised sentence-level summary relevance prediction with unsupervised clustering-based document-level score enhancement. It utilises a Legal BERT-based Multi-Layer Perceptron model to estimate the summary relevance of each sentence, then refines these scores through deep embedded sentence clustering to incorporate the document's global context.

9 Legal Named Entity Recognition

NER identifies and categorises textual mentions into predefined types such as organisations, persons and locations [70]. In the legal domain, NER extends to specialised recognition tasks that focus on extracting entities unique to legal texts, such as laws, legal norms, and procedural terms. This specialised form of NER is crucial for structuring legal documents and enhancing legal IR systems. Unlike general NER systems that handle common entity types, legal NER must navigate the complex language and structured format of legal documents, motivating the need for systems and methodologies specifically tailored to the legal context.

9.1 Datasets

Leitner et al. [72] release a German legal NER corpus drawn from federal court decisions, comprising about 67,000 sentences and more than two million tokens. It contains roughly 54,000 manually annotated entities across 19 fine-grained, domain-specific classes, such as court, judge, lawyer, law, person and legal literature, alongside over 35,000 TimeML-based [109] time expressions. The annotations cover both broad categories, such as location, person and organisation and more specialised ones, such as legal norms and case-by-case regulations, distinguishing between different types of legal acts and literature. Annotation proceeded through multiple iterations to refine guidelines and ensure high-quality labels.

Păis et al. [110] introduce the LegalNERo corpus, a manually annotated resource for NER in the Romanian legal domain, featuring 370 legal documents annotated with five entity types: person, location, organisation, time expressions and legal references. This corpus was developed to support both specific legal domain NER tasks and more general NER applications by enabling compatibility with existing general-purpose NER systems. The corpus includes rich entity annotations, with legal references showing the highest token count per entity, indicating their complexity and length. The annotation workflow involved several refinement cycles, inter-annotator agreement measured by Cohen's kappa and conversion of entities to RDF, ensuring accuracy and usefulness for legal NER research.

Au et al. [7] present the E-NER dataset, an annotated collection derived from the US Securities and Exchange Commission's EDGAR filings, designed for legal NER. This dataset contains filings that are rich in text, such as quarterly reports and significant event announcements, from which sentences were extracted and annotated with seven named entity classes more tailored to legal content than those in the standard CoNLL dataset [131]. The entities include person, location, organisation, government, court, business and legislation/act, adjusting the CoNLL classes to better suit legal documents. E-NER contains longer sentences compared to CoNLL and includes detailed annotations of financial entities from legal company filings.

Kalamkar et al. [65] release a legal NER corpus with 46,545 entities of 14 types extracted from Indian High Court and Supreme Court judgements. The corpus is divided into preamble and judgement sections and covers entities such as court, petitioner, respondent, and statute. The training set, drawn from judgements between 1950 and 2017, contains 29,964 entities, while the development and test sets cover cases from 2018 to 2022, ensuring no training leakage. This dataset not only facilitates training and evaluation of NER models specific to the legal domain but also provides a structured framework for assessing the performance of NER systems on legal texts. Their approach leverages a combination of manual annotation and ML techniques to ensure the precision of entity recognition in legal judgements.

9.2 Approaches

Dozier et al. [35] conduct early research on NER in legal texts, including US case law and pleadings, by combining lookup methods, contextual rules and statistical models to identify entities such as judges, attorneys and legal terms. Their system adapts these approaches to the specialised context of legal texts, processing various types of documents and extracting legal entities. This work highlights the challenges and necessary adaptations for deploying NER in the legal domain, where the specialised language and high accuracy are required for successful legal analysis.

Păis et al. [110] develop a legal NER model that combines Bi-LSTM layers with a Conditional Random Field (CRF) output layer and leverages multiple data sources and embedding types. The architecture integrates pre-trained word embeddings, character embeddings, and gazetteer entries from GeoNames[7] and JRC-Names [127], along with known legal affixes, to enrich text representations. During training, word embeddings are fine-tuned while character embeddings are learned dynamically through the Bi-LSTM layers, improving generalisation to unseen texts. Built on a modified version of NeuroNER [30], the system supports online serving and employs dropout for regularisation and gradient clipping to mitigate exploding gradients. The authors also explore ensembles of different model configurations, evaluating performance via precision, recall and F1 scores against a gold-standard corpus.

Smădu et al. [123] explore domain adaptation in legal NER, focusing on the Romanian and German languages. Their architecture combines a pre-trained BERT layer for feature extraction with Bi-LSTM networks to handle sequence dependencies and CRFs for sequence tagging. Their approach employs domain adaptation techniques through a gradient reversal layer connected to a domain discriminator, aimed at reducing domain-specific biases and enhancing feature transferability across domains. The model is trained on both legal and general corpora through adversarial learning, aiming to improve transferability across domains. Results show marginal gains for German but a performance drop on the Romanian legal dataset, indicating that benefits vary by language and domain.

Adhikary et al. [1] publish LeDa, a legal data annotation system designed to address the challenges of extracting legal entities and concepts from case documents. Unlike traditional sequence labelling tools, LeDA enables annotators to dynamically define new legal concepts during annotation. The system also incorporates a meta-annotation mechanism for adjudicating conflicting annotations.

10 Legal Argument Mining

LAM applies NLP to identify and extract arguments from legal documents, automating the detection of claims, premises, and their interrelations to enhance legal research and practice. By reconstructing both the local structure of individual arguments and the global network of relations between them, it supports legal reasoning, exposing chains of reasoning that inform judicial decisions and support tasks such as conflict resolution and why-question answering [101]. To meet these demands, recent work focuses on domain-specific annotation schemes and

[7] https://www.geonames.org/

advanced DL models capable of handling the intricacies of legal language and argument structure. Effective argument mining therefore relies on identifying elementary argumentative units, modelling their rhetorical relations and determining whether these structures can be derived automatically.

10.1 Datasets

Poudyal et al. [108] present the ECHR corpus for LAM, a structured dataset of 42 decisions from ECHR, annotated with argumentative components: premises, conclusions and non-argumentative text. The corpus facilitates research on argument mining by enabling three key tasks: argument clause recognition, clause relation prediction and premise/conclusion classification. The annotation process involved legal experts, with iterative refinements leading to 80% inter-annotator agreement. Habernal et al. [55] develop an annotation scheme for ECHR judgements, grounded in legal argumentation theory and designed to capture argument spans such as claims, premises and their interrelations. Using this scheme, the authors compile and manually annotate a large corpus of 373 ECHR decisions, comprising approximately 2.3 million tokens and over 15,000 argument spans. Six law students performed the annotation under expert supervision and achieved high inter-rater agreement (as measured via Krippendorff's alpha, with values close to 0.80).

Grundler et al. [54] introduce Demosthenes, a corpus of 40 Court of Justice of the European Union (CJEU) decisions on fiscal state aid, annotated at the sentence level for argument mining. Each decision (written in English and spanning 2000–2018) was obtained from EUR-Lex and manually annotated to capture its argumentative reasoning, focusing on the "Findings of the Court" section where the judgement's legal arguments are laid out. Using an iteratively refined annotation guideline, two experts with legal domain expertise labelled each sentence in this section as an argumentative element (premise or conclusion), further denoting each premise as legal or factual and assigned each argument to a category in a legal argumentation scheme typology.

10.2 Approaches

Palau and Moens [101] present pioneering research in LAM, focusing on detecting, classifying and structuring arguments within legal texts. Their work introduces methods to automatically identify arguments, distinguish argumentative from non-argumentative sentences and classify argumentative propositions as either premises or conclusions using statistical classifiers such as maximum entropy models, Naive Bayes classifiers and SVM. The authors utilise the Araucaria and the ECHR corpora for evaluation. Results demonstrate that statistical methods can distinguish argumentative sentences, achieving approximately 80% accuracy on the ECHR corpus. Additionally, they discuss methods to resolve argument segmentation challenges through structural and semantic analyses, proposing that semantic relatedness measures (based on ontology or corpus-derived statistics) can enhance argument boundary detection. Finally, they investigate detecting argument structures through rhetorical pattern analysis and suggest employing context-free grammars as an initial step toward full argumentative parsing.

Zhang et al. [148] propose a graph-based framework for LAM that replaces the traditional pipeline approach with an end-to-end architecture. By modelling each legal document as a graph – where nodes represent text segments and edges capture sequential or semantic relationships – they mitigate error propagation across sub-tasks. Their method employs virtual node graph augmentation, which adds a global node connected to all text segments and a collective classification algorithm that iteratively refines predictions using neighbouring node labels. They evaluate on the ECHR and the CJEU datasets using Graph Convolutional Network (GCN) and Residual Gated GCN (ResGCN) models. In particular, ResGCN demonstrates better performance on both datasets, surpassing baseline methods in classifying premises, conclusions and non-argumentative text.

Santin et al. [115] propose a novel annotation scheme for predicting argument structures in CJEU fiscal state aid decisions, addressing the scarcity of annotated resources and the complexity of legal reasoning. Building

Table 4. Overview of datasets and their usage across different legal NLP tasks and cited studies.

Task	Datasets	Identified Studies
LJP	CAIL datasets by Xiao et al. [141]	Feng et al. [41], Tong et al. [132], Xu et al. [142], Yang et al. [143, 143], Zhang et al. [149], Zhong et al. [152, 153]
LTC	LEDGAR by Tuggener et al. [133] EURLEX57K by Chalkidis et al. [18]	Mamooler et al. [86] Song et al. [124]
NER	LegalNERo by Păis et al. [110] German LER by Leitner et al. [72]	Smădu et al. [123]
LAM	ECHR by Poudyal et al. [108] Demosthenes by Santin et al. [115]	Zhang et al. [148] Santin et al. [115], Zhang et al. [148]

on the Demosthenes corpus, they refine their previous dataset by distinguishing a richer set of inferential relations, including direct support, indirect support (support from failure), rebuttal, undercut and rephrase links, which captures the logical and discursive connections support judicial arguments. Using the extended dataset, they conducted an empirical study comparing DistilRoBERTa [113] with an ensemble of attentive residual networks [44] (ResAttArg) for link prediction, exploring variations in training (with/without oversampling and different link distance thresholds). The ResAttArg ensemble outperforms the distilled transformer and does so with lower computational demand.

Habernal et al. [55] utilise advanced NLP techniques to mine legal arguments from ECHR decisions. Specifically, they employ a multi-task transformer-based model, which is a DL approach designed for both argument identification and classification. This model leverages the capabilities of transformers to process and understand complex legal texts, outperforming previous legal NLP models according to expert evaluations.

Table 4 maps each dataset to every cited study that employs it, covering both the current task and other legal NLP tasks discussed in the survey.

11 Large Legal Datasets

Training LLMs for legal NLP requires extensive legal corpora that are transparent in sourcing, safeguard privacy and minimise bias to ensure fairness and accuracy. These corpora serve as the foundation for developing models capable of handling diverse legal tasks. To assess model performance, evaluation benchmarks provide structured datasets and standardised metrics for tasks such as judgement prediction, QA, case retrieval and entailment. Together, large legal corpora and evaluation benchmarks can support the advancement of reliable legal AI applications.

11.1 Evaluation benchmarks

Zheng et al. [151] introduce CaseHOLD, a novel benchmark for evaluating NLP models in the legal domain, designed to address the challenge of identifying the legal holdings from case texts. The dataset contains over 53,000 multiple-choice questions derived from the US case law citations, where each question requires the identification of the correct holding from a set of potential answers. This task, simulating a fundamental skill taught in law school, involves contextual understanding and application of legal rules to factual situations. CaseHOLD is aimed at enhancing model training by focusing on semantic matching and the ability to discern legal principles. The dataset is structured to provide a challenging yet accessible resource for NLP researchers, with a clear focus on promoting deeper understanding and application of legal rules in automated systems. Each item presents a cited

passage as a prompt followed by one correct holding and four closely related incorrect options, encouraging models to demonstrate genuine legal reasoning.

Chalkidis et al. [21] introduce the Legal General Language Understanding Evaluation (LexGLUE) benchmark, a comprehensive suite of datasets aimed at assessing the capabilities of NLP models across various legal tasks. The benchmark covers datasets, such as ECHR [17], SCOTUS,[8] EUR-Lex, LEDGAR [133], UNFAIR-ToS [77] and CaseHOLD [151], each chosen for its complexity, relevance, and need for legal expertise. These datasets cover a range of tasks from multi-label and multi-class classification to multiple-choice questions and are split chronologically into training, development and test sets to provide standardised evaluation metrics. For instance, ECHR datasets focus on violations of European Convention of Human Rights provisions; theSCOTUS database classifies the US Supreme Court opinions by legal issues; the EUR-Lex database involves labelling EU laws with EuroVoc concepts; LEDGAR classifies provisions of the US contracts; UNFAIR-ToS identifies unfair terms in online service agreements; and CaseHOLD involves QA about legal rulings.

Chalkidis et al. [22] introduce FairLex, a multilingual benchmark suite of four legal datasets: ECHR [17], SCOTUS, FSCS and CAIL. The suite evaluates fairness in NLP across several jurisdictions (Europe, the United States, Switzerland, and China) and five languages (English, German, French, Italian, and Chinese). Each dataset is aligned with a specific legal task: ECHR violation prediction for ECHR; issue-area classification for SCOTUS; case-approval prediction for FSCS; and crime-severity prediction for CAIL. All datasets are chronologically split into training, development and test sets. FairLex supports demographic, regional and topical fairness analysis by recording sensitive attributes, including defendant state in ECHR, decision direction in SCOTUS, legal area in FSCS, and gender and region of origin in CAIL.

Rabelo et al. [111] summarise the 8th Competition on Legal Information Extraction and Entailment (COLIEE 2021), which featured five tasks across case and statute law, engaging participants from various teams to apply diverse NLP approaches. The competition tasks included case law retrieval and entailment, as well as statute law retrieval and entailment with and without prior retrieved data. Specifically, task 1 focused on extracting relevant supporting cases from a corpus, while task 2 involved identifying paragraphs from cases that entail a given new case fragment. For statute law, tasks 3 and 4 entailed retrieving and answering questions based on civil code statutes, with task 5 challenging participants to answer without pre-retrieved statutes. The datasets used varied in complexity, from 4,415 case files in task 1 with a need to identify noticed cases without relying on citations, to the civil code-based tasks 3, 4 and 5 which adapted to recent legal revisions in Japanese law and excluded untranslated parts, reflecting the ongoing evolution and challenge in legal NLP applications.

Barale et al. [9] present AsyLex, a pioneering dataset tailored for refugee law applications, featuring 59,112 documents from Canadian refugee status determinations spanning from 1996 to 2022. This dataset is designed to enhance the capabilities of NLP models in legal research by providing 19,115 gold-standard human-annotated and 30,944 inferred labels for entity extraction and LJP. Key contributions include anonymizing decision documents, employing a robust annotation methodology and creating datasets for specific NLP tasks, such as entity extraction and judgement prediction.

Niklaus et al. [97] present LEXTREME, a multilingual benchmark for evaluating LMs on legal NLP tasks. Drawing on legal NLP research published between 2010 and 2022, they curate 11 human-annotated datasets spanning 24 languages and multiple legal domains. To ensure fair comparison across models, the authors introduce two aggregate metrics—the dataset aggregate score and the language aggregate score—and show that performance on LEXTREME rises with model size. The benchmark covers three task types: single-label text classification, multi-label text classification and NER. Where possible, it preserves existing train, development and test splits, otherwise creating random splits.

[8]https://www.supremecourt.gov

Park and James [104] explore the creation of a Natural Language Inference (NLI) dataset within the legal domain, focusing on criminal court verdicts in Korean. Their methodology includes the innovative use of adversarial hypothesis generation to challenge annotators and enhance the robustness of the dataset, supported by visual tools for hypothesis network construction. The data collection involves extracting context from verdicts and augmenting it using Easy Data Augmentation [138] techniques and round-trip translation to generate a dataset for training and testing NLI models. The study highlights issues such as annotators' limited domain knowledge and challenges in handling long contexts but provides solutions, such as targeted data collection and the use of gamification to boost annotator engagement and productivity.

Goebel et al. [51] summarise COLIEE 2023, featuring four tasks across case and statute law with participation from ten different teams engaging in multiple tasks. task 1 involves legal case retrieval, requiring participants to extract supporting cases from a corpus and task 2 focuses on legal case entailment, identifying paragraphs that entail aspects of a new case. Task 3 and 4, based on Japanese civil code statutes from the bar exam, involve retrieving relevant articles and verifying statements, respectively. The competition leverages a dataset of over 5,700 case law files and introduces new query cases and test questions sourced from recent bar exams, testing the efficacy of different teams' approaches in handling complex legal texts and hypotheses in a controlled competitive environment.

Östling et al. [157] introduce the Cambridge Law Corpus (CLC), a legal dataset featuring 258,146 cases from UK courts, dating from the 16th century to the present. The corpus includes raw text and metadata across various court types and is structured in XML format for ease of use and annotated for case outcomes in a subset of 638 cases. Additionally, the CLC is supported by a Python library for data manipulation and ML applications.

11.2 Large corpora for pre-training

Henderson et al. [56] introduce the "Pile of Law," the first at-scale legal text collection, containing a 256 GB dataset of open-source English-language legal and administrative data. This dataset includes contracts, court opinions, legislative records, and administrative rules, curated to explore data sanitation norms across legal and administrative settings and serve as a tool for pre-training legal LLMs. They emphasise the legal norms governing privacy and toxicity filtering, detailing how the dataset reflects these norms through built-in filtering mechanisms in the collected data, which include court filings, legal analyses and government publications. By analysing how legal and administrative entities handle sensitive information and potentially offensive content, the paper provides actionable insights for researchers to improve content filtering practices before pre-training LLMs, thereby enhancing the ethical use of NLP in legal applications.

Niklaus et al. [98] present the MultiLegalPile, the largest open-source multilingual legal corpus available, totalling 689 GB and spanning 17 jurisdictions across 24 languages. This extensive dataset is designed to facilitate training of LLMs within the legal domain, featuring diverse legal text types including case law, legislation and contracts, predominantly in English due to the integration of the "Pile of Law" [56] dataset. Through careful regex-based filtering from the mC4 corpus and manual reviews, the team ensures high precision in legal content selection.

12 Legal Language Models and Methods for Legal Domain Adaptation

In the fast-moving field of NLP, LLMs have become a key tool for processing and understanding large amounts of unstructured text data. These models, initially trained on broad datasets such as Wikipedia, have shown great skill across various language tasks. Building on this success, the legal technology community is increasingly interested in using these powerful models for legal NLP applications. This involves adapting these general-domain models to legal texts and further training them on specialised legal documents. Such efforts aim to reduce the domain gap and customise the models to better understand the complex language used in legal documents. In this section,

we will explore how these so-called "foundation" models are being adapted and applied within the legal domain to enhance legal NLP applications.

Following the methodology of this survey, this section studies all peer-reviewed LMs or related methods. However, due to the challenges present in the legal domain, there are many legal LMs that have not undergone peer review. Given the scarcity of adequate peer-reviewed resources, our research has focused on the investigation of, in order of priority, the peer-reviewed sources, then the most well-known and widely used non-peer-reviewed legal LMs. Despite their lack of formal peer review, these models have gained considerable attention and usage in the field, and some are expected to be published in peer-reviewed venues in the future.

12.1 Language Models

Chalkidis et al. [20] present an in-depth analysis of applying BERT in the legal domain, showcasing the need for domain-specific adaptation to enhance performance on legal NLP tasks. They explore three strategies: using standard BERT directly, further pre-training a BERT model on legal corpora, and pre-training from scratch with legal-specific data. Their study found that both further pre-training and pre-training from scratch generally outperform the use of BERT directly. They introduce legal-bert, which includes versions for varied computational capacities and demonstrates competitive performance with a lower environmental impact.

Xiao et al. [140] introduce Lawformer, a Longformer-based [10] LM adapted for Chinese legal texts, designed to handle extensive document lengths common in legal data. Recognising the limitation of standard PLMs with shorter token capacities, Lawformer employs a unique combination of sliding window, dilated sliding window, and global attention mechanisms to process long texts, making it suitable for legal AI tasks, such as LJP and LQA. Pre-trained on a vast corpus of Chinese legal documents segmented into criminal and civil cases, Lawformer integrates complex sequential dependencies across tokens using these attention techniques, enhancing model performance for legal-specific tasks.

In the development of specialised NLP tools for Arabic legal texts, a model specifically tailored to the unique linguistic features of Arabic jurisprudence was designed called AraLegal-BERT [2]. This model enhances NLP applications within the legal field by adapting BERT technology to Arabic's specific content needs, involving pre-training BERT from scratch using a broad range of legal documents, including legislative materials and contracts.

Colombo et al. [27] introduce SaulLM-7B, a novel LLM specifically designed for legal text comprehension and generation, built on the 7 billion parameter Mistral [62] architecture. This model is trained on an extensive English legal corpus, designed to meet the unique challenges of legal syntax and terms. SaulLM-7B uses a two-tier training approach: continued pre-training on a carefully curated 30 billion token legal dataset and an innovative instruction fine-tuning method, incorporating both generic and legal-specific instructions to enhance the model's performance on legal tasks.

Shi et al. [122] develop Legal-LM, a specialised LM tailored for Chinese legal consulting, enhanced with a KG to address domain-specific challenges such as data veracity and non-expert user interaction. The framework involves several steps: extensive pre-training on a rich corpus of legal texts integrated with a legal KG, keyword extraction and Direct Preference Optimisation to refine responses and the use of an external legal knowledge base for data retrieval and response validation. This multi-faceted approach ensures that Legal-LM not only comprehends complex legal language but also generates precise and user-aligned legal advice.

12.2 Methods for Improving In-Domain Adaptability of Legal Language Models

Li et al. [74] explore a novel adaptation of LMs for the legal domain by integrating domain-specific unsupervised data from public legal forums to optimise prefix domain adaptation, a parameter-efficient learning approach that trains only about 0.1% of the model's parameters. They introduce a training methodology where a deep prompt

is specifically tuned using a domain-adapted prefix from legal forums and then utilised in various legal tasks, demonstrating improved few-shot performance compared to full model tuning methods, such as LEGAL-BERT [20]. This approach reduces computational overhead while maintaining or exceeding performance metrics across multiple legal tasks, suggesting an efficient and scalable model for legal NLP applications.

Mamakas et al. [85] explore strategies for adapting pre-trained transformers to cope with the challenges of long legal texts within the LexGLUE benchmark, focusing on extending input capabilities and enhancing efficiency. They modify Longformer [10], originally extending up to 4,096 subwords, to process up to 8,192 subwords by reducing local attention window size and incorporating a global token at the end of each paragraph to facilitate information flow across longer texts. Additionally, they adapt legal-bert to employ TF-IDF representations to manage longer documents, introducing variants, such as TF-IDF-SRT-LegalBERT, which deduplicates and sorts subwords by TF-IDF scores; and TF-IDF-EMB-LegalBERT, which incorporates a TF-IDF embedding layer. These adaptations aim to combine the robust capabilities of transformers with the practical requirements of handling extensive legal documents, surpassing the performance of traditional linear classifiers while maintaining computational efficiency.

13 Open Research Challenges

Despite researchers' efforts in the this interdisciplinary field and extensive advancements in AI techniques, a number of *open research challenges* (ORCs) still exist. In this section, we identify the key ORCs, and provide advice and directions for future work to overcome these challenges.

ORC1: Bias and Fairness. Bias and fairness are crucial concerns in the field of AI, especially at the intersection with the legal domain where decisions can deeply impact individuals' lives. The scarcity of unbiased data in legal domains such as case law complicates the training of AI models, as these models often learn from historical decisions that may reflect existing human biases [36, 130]. This reliance on biased datasets can, in turn, lead to unfair and biased outcomes in downstream classification and prediction tasks. Addressing these issues is critical to ensure that AI-driven legal decisions uphold the standards of impartiality and fairness required for justice.

ORC2: Privacy Concerns. Privacy concerns in legal NLP are critical, as models often handle highly sensitive documents such as court records [145]. Beyond basic anonymization, advanced techniques – such as differential privacy – add statistical noise during training to ensure that individual data points have minimal influence on the model's output. Adversarial training further bolsters privacy by simulating potential attacks and guiding the model to suppress identifiable attributes in its representations. Privacy-preserving fine-tuning methods – such as applying differential privacy during fine-tuning, federated learning or knowledge distillation – allow models to adapt to new legal tasks without exposing sensitive data. However, implementing these techniques requires careful calibration, as excessive privacy constraints may degrade model accuracy. This challenge motivates the need for continued research into privacy-preserving NLP methods tailored to the unique demands of the legal domain.

ORC3: Interpretability and Explainability. The ability to interpret and explain the outputs of AI-powered systems is of critical importance across various applications in legal NLP, yet these aspects remain underexplored in many contexts. Relating to ORC1, the ability to trace and comprehend the decision-making process of AI systems is essential for identifying and mitigating biases. Transparent and understandable AI systems help build trust and ensure they are used responsibly, which is particularly important in legal contexts where decisions significantly impact people's lives. Improving these aspects of AI models is necessary to their ethical use, ensuring they meet the high standards of fairness required in legal proceedings. In turn, this will promote wider adaption, allowing the promises of these systems – increased productivity, fairness, and reduced costs – to be realised in practice.

ORC4: Annotation Process and Transparency. Annotation quality and transparency are critical challenges in legal NLP. A growing discussion in the community focuses on descriptive versus prescriptive annotation [112]. Descriptive annotation captures the full spectrum of annotator subjectivity, preserving diverse interpretations of legal texts and proving valuable for interpretative tasks such as LQA or contract analysis. In contrast, prescriptive annotation enforces a single, consistent standard that is essential for tasks requiring strict adherence to legal norms, such as LJP or statute classification. The choice between these paradigms depends on the specific task, domain context and intended downstream application. Equally, transparent annotation practices are rarely documented in legal NLP studies, with many reports omitting details about annotators' backgrounds and the procedures employed. Comprehensive documentation of the annotation process is important for assessing dataset quality, ensuring fairness and building trust in legal NLP systems. Ultimately, explicit decisions regarding both the annotation paradigm and transparency measures are necessary to produce reliable, suitable datasets.

ORC5: Scarcity of Reliable Annotated data. One of the key ORCs in legal NLP is the scarcity of annotated data, which limits the development of robust models due to the high cost and expertise required for labelling complex legal texts. Data augmentation can offer a promising approach to mitigate this issue, yet it introduces its own set of challenges. The complexity of legal texts complicates the application of general augmentation techniques. Ensuring the quality of augmented data remains critical, with efforts focusing on preserving semantic and syntactic integrity to prevent models from learning incorrect patterns [105]. Additionally, the effectiveness of augmentation varies across models, necessitating tailored strategies, while scalability concerns arise with computationally intensive methods [49]. The predominant focus on English limits applicability to multilingual legal contexts and in extremely low-resource scenarios, there is a risk of overfitting to augmented data [49]. Future research should prioritise developing domain-specific augmentation techniques that maintain the nuances of legal text, explore flexible approaches, enhance scalability, and extend support diverse languages, thereby addressing these persistent challenges in legal NLP.

ORC6: Multilingual Capabilities. In legal NLP, enhancing multilingual capabilities remains an underdeveloped area. While efforts, such as MultiLegalPile [98] have begun to address this, there remains a gap in research for many languages, including, but not limited to, Persian and Arabic. These limitations restrict the application of legal NLP across diverse legal systems worldwide, hampering broader adoption and accessibility. Multilingual capabilities introduce unique challenges for legal NLP models, primarily due to the distinct linguistic structures of each language, which often require extensive fine-tuning to ensure accuracy and relevancy in legal contexts. Furthermore, each legal system possesses its own set of terms and document standards, which can vary dramatically from one language to another. Therefore, expanding research into these and other underserved languages is essential for making NLP tools universally applicable.

ORC7: Ontologies and Knowledge Graphs. The use of ontologies in the legal domain is relatively sparse, yet it holds considerable potential to enhance the robustness of AI methodologies. Ontologies or knowledge graphs can also enable AI models to draw accurate inferences regarding the relationships between entities, thereby improving reasoning and decision making over complicated legal texts. However, utilising ontologies in legal NLP faces unique challenges. The complexity of legal language and the concept of "open texture," where the meaning of legal terms can evolve over time, can complicate the creation of static ontological models [92]. Legal ontologies must be dynamic, reflecting changes in law and its interpretation over time. Additionally, the integration of real-world and legal concepts within ontologies presents further complexity, as it requires accommodating both legal terms and their relevant real-world contexts [92].

ORC8: Pre-processing Legal Text. Pre-processing legal texts is challenging because legal documents often consist of raw texts that requires extensive cleaning and transformation before use in ML models. Their complex, nested structures, such as clauses within clauses and cross-references to other cases, statutes or provisions, make it hard

to segment them into coherent units for analysis. Without addressing these complexities, fine-tuning LMs on raw legal data becomes impractical, limiting the performance of legal NLP applications.

ORC9: Reinforcement Learning from Human Feedback (RLHF). The use of RLHF within the legal domain is notably scarce. Currently, there is only one peer-reviewed work [94] available that explores this approach. This indicates an opportunity for research and development in this area, as RLHF could potentially enhance the capability of NLP models to learn and to make decisions based on complex legal data under human guidance. Further exploration into this method could lead to more responsive and adaptable legal NLP systems. However, due to the complex nature of legal reasoning and the need for accurate legal knowledge in the human feedback phase, integrating RLHF into legal NLP pipelines poses some challenges. In particular, legal experts such as lawyers and judges must provide guidance to ensure the AI models accurately interpret and apply complex legal concepts, which is a barrier due to the significant cost of employing such professionals. Nonetheless, if RLHF can demonstrate further downstream savings due to improved worker efficiency, the up-front cost of employing legal professionals for data annotation tasks may be a worthwhile investment.

ORC10: Expanding Legal Domain Coverage. There is a noticeable gap in the research across various areas of the legal domain, including Intellectual Property, Criminal Law, Banking Law, Family Law, and Human Rights Law. These fields have seen limited exploration across all legal NLP tasks, such as LQA and other applications. Expanding research into these areas is essential for developing comprehensive automated legal systems that can provide tailored solutions and insights highly relevant to these sectors of law.

ORC11: Small Language Models (SLMs). Research into SLMs specific to the legal domain is notably absent. Although some work explored notions of distillation and the use of smaller transformers, there has not been a significant exploration into the efficiency considerations of employing large language models in the legal domain. Addressing this gap could lead to more efficient, resource-conscious solutions that still maintain high performance in legal text processing and analysis. The development of SLMs tailored for legal applications could revolutionise the accessibility and scalability of legal NLP tools.

ORC12: Domain-Specific Efficient Fine-Tuning. Domain-specific efficient fine-tuning within the legal field remains underexplored, with only two known studies addressing it [74, 82]. Legal texts feature complex structures and specialised vocabulary that standard LLMs may not capture without substantial adaptation. Moreover, the legal domain covers a vast array of document types, such as case law, statutes and contracts, each requiring tailored model strategies. This diversity makes it imperative to develop fine-tuning methods that not only adapt a model generally but also tailor it to the specific characteristics of each document type. Most existing approaches fine-tune the entire model, which can be resource intensive. More focused research could enable efficient fine-tuning of legal LLMs using fewer resources and improve their practical deployment.

ORC13: Legal Logical Reasoning. Complex legal logical reasoning remains a challenge in LJP, particularly in predicting prison terms. Current SOTA methods struggle to achieve high accuracy in this area, highlighting a clear need for improved approaches that better handle legal reasoning. Multi-hop QA Models with knowledge graphs, or reasoning LMs, may be an avenue worth investigating here.

ORC14: Legal Named Entity Recognition. Legal NER focuses on specific challenges such as disambiguating titles, resolving nested entities, handling co-references and processing PDFs that are not machine-readable. Despite its crucial role in structuring and interpreting legal documents, research on legal NER remains limited, as shown in Figure 1.

ORC15: Stochastic Parrots. The concept of "Stochastic Parrots" pertains particularly to LLMs. It shows the concern that these models often do not truly understand language but merely mimic human patterns. This mimicry can

Table 5. Summary of existing ORCs in each area. A direct relationship is denoted with a check-mark (✓).

Open Research Challenges	LQA	LJP	LTC	LDS	NER	AM	LLMs	Corpora
Bias and Fairness	✓	✓	✓	–	–	–	✓	✓
Privacy Concern	✓	✓	✓	✓	✓	–	✓	✓
Interpretability and Explainability	✓	✓	✓	–	–	✓	✓	–
Annotation Process and Transparency	✓	✓	✓	✓	✓	✓	✓	✓
Scarcity of Reliable Annotated data	✓	✓	✓	✓	✓	✓	✓	✓
Multilingual Capabilities	✓	✓	✓	✓	✓	✓	✓	✓
Ontology	✓	✓	✓	✓	✓	✓	✓	–
Pre-processing Legal Text	✓	✓	✓	✓	✓	✓	✓	✓
RLHF	✓	✓	✓	✓	–	✓	✓	–
Expanding Legal Domain Coverage	✓	✓	✓	✓	✓	✓	✓	✓
SLMs	–	–	–	–	–	–	✓	–
Domain-Specific Efficient Fine-Tuning	✓	✓	✓	✓	–	–	✓	–
Legal Logical Reasoning	✓	✓	✓	–	–	✓	✓	–
Legal NER	–	–	–	–	✓	✓	–	–
Stochastic Parrots	–	–	–	–	–	–	✓	–
RAG	✓	✓	–	–	–	✓	✓	–

lead to unreliable outcomes, especially in critical legal situations, if the models are not trained on high-quality, unbiased datasets. The risk is notably significant in LJP, where training on biased or unfair data could lead to irreversible outcomes, as discussed in Bender et al. [11]'s work on the limitations of LLMs. Ultimately, we must ensure that tools and algorithms aimed at automating legal processes do not become *decision makers*, rather that they are framed as *decision support* tools. More research is required to fully grasp the sociotechnical aspects of NLP in the legal domain.

ORC16: Retrieval-Augmented Generation. The legal domain presents particular challenges because documents are lengthy, contain numerous cross-references, and exhibit complex linguistic structures. These characteristics can cause LLMs to hallucinate when asked for precise answers. RAG offers a promising remedy by incorporating relevant passages directly into the generation process, overcoming LLM input-length limits and improving both relevance and contextual accuracy. However, applying RAG in legal settings still poses challenges: handling documents from multiple jurisdictions, ensuring that retrieved material remains temporally current, addressing multilingual texts, and mitigating retrieval biases are all significant problems that must be resolved.

Summary. Table 5 illustrates the connections between ORCs and the discussed research areas. As shown, most ORCs are related to LJP, LQA, LTC and LLMs, indicating more extensive research in these areas. Nonetheless, it is evident that there remains significant work to be done in all areas of focus, and this work is likely to be of a cross-disciplinary nature, spanning a number of research communities.

14 Conclusion

Advances in AI and NLP have improved legal NLP techniques and models, reducing the difficulty of engaging in legal processes for laypersons, and easing workloads and manual labour for professionals. This survey provides a comprehensive overview of the advancements in NLP techniques used in the legal domain, paying special attention to the unique characteristics of legal documents. We also reviewed existing datasets and LLMs tailored for the legal domain. Legal NER research spans multiple languages and utilises diverse methods, from rule-based to BERT-based models. LDS has largely focused on extractive and abstractive methods, ranging from TF-IDF to transformer-based models. LAM now automates the detection of claims, premises, and their links through

domain-specific annotation schemes and graph-based or residual-network approaches, supporting legal reasoning tasks such as conflict resolution. In LTC, multi-class classification dominates, with DL architectures, such as CNNs and Bi-LSTMs widely used. LJP primarily focuses on Chinese datasets with DL approaches, such as CNNs. LQA often leverages IR techniques such as BM25, with a significant focus on statutory law. Finally, we explored key ORCs, such as the need for domain-specific fine-tuning strategies, addressing bias and fairness in legal datasets and the importance of interpretability and explainability. Other challenges include the development of more robust pre-processing techniques, handling multilingual capabilities and integrating ontology-based methods for more accurate legal reasoning.

Acknowledgments

This work is partially supported by an Australian Research Council (ARC) Future Fellowship Project (Grant No. FT240100022) and the Swiss National Science Foundation (SNSF) under contract number CRSII5_205975.

References

[1] Subinay Adhikary, Dwaipayan Roy, Debasis Ganguly, Shouvik Kumar Guha, and Kripabandhu Ghosh. 2023. Leda: a system for legal data annotation. In *Legal Knowledge and Information Systems*. IOS Press, 367–370.

[2] Muhammad Al-qurishi, Sarah Alqaseemi, and Riad Souissi. 2022. AraLegal-BERT: A pretrained language model for Arabic Legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*.

[3] Intisar Almuslim and Diana Inkpen. 2022. Legal Judgment Prediction for Canadian Appeal Cases. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*. 163–168.

[4] Dang Hoang Anh, Dinh-Truong Do, Vu Tran, and Nguyen Le Minh. 2023. The Impact of Large Language Modeling on Natural Language Processing in Legal Texts: A Comprehensive Survey. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*.

[5] Arian Askari, Suzan Verberne, and Gabriella Pasi. 2022. Expert Finding in Legal Community Question Answering. In *Advances in Information Retrieval: 44th European Conference on IR Research (ECIR 2022)*. 22–30.

[6] Arian Askari, Zihui Yang, Zhaochun Ren, and Suzan Verberne. 2024. Answer Retrieval in Legal Community Question Answering. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval (ECIR 2024)*. 477–485.

[7] Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — An Annotated Named Entity Recognition Corpus of Legal Text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 246–255.

[8] Purbid Bambroo and Aditi Awasthi. 2021. LegalDB: Long DistilBERT for Legal Document Classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. 1–4.

[9] Claire Barale, Mark Klaisoongnoen, Pasquale Minervini, Michael Rovatsos, and Nehal Bhuta. 2023. AsyLex: A Dataset for Legal Language Processing of Refugee Claims. In *Proceedings of the Natural Legal Language Processing Workshop 2023*. 244–257.

[10] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[12] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. *Advances in Information Retrieval* (2019), 413–428.

[13] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2023. DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law* (2023), 1–38.

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Article 159.

[15] Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in German civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2015–2027.

[16] Ilias Chalkidis. 2023. ChatGPT may Pass the Bar Exam soon, but has a Long Way to Go for the LexGLUE benchmark. arXiv:2304.12202

[17] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4317–4323.

[18] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*. 78–87.

[19] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6974–6996.

[20] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

[21] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4310–4330.

[22] Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

[23] Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. EQUALS: A Real-world Dataset for Legal Question Answering via Reading Chinese Laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*. 71–80.

[24] Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multi-Task Learning in Natural Language Processing: An Overview. *ACM Comput. Surv.* 56, 12 (2024).

[25] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. arXiv:2405.01769

[26] Odysseas S. Chlapanis, Ion Androutsopoulos, and Dimitrios Galanis. 2024. Archimedes-AUEB at SemEval-2024 Task 5: LLM explains Civil Procedure. arXiv:2405.08502

[27] Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. SaulLM-7B: A pioneering Large Language Model for Law. arXiv:2403.03883

[28] Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges. *IEEE Access* 11 (2023), 102050–102071.

[29] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (2024), 64–93.

[30] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 97–102.

[31] Aniket Deroy and Subhankar Maity. 2023. Questioning Biases in Case Judgment Summaries: Legal Datasets or Large Language Models? arXiv:2312.00554

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[33] João Dias, Pedro A. Santos, Nuno Cordeiro, Ana Antunes, Bruno Martins, Jorge Baptista, and Carlos Gonçalves. 2022. State of the Art in Artificial Intelligence applied to the Legal Domain. arXiv:2204.07047

[34] Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-Aware Unsupervised Summarization for Long Scientific Documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1089–1102.

[35] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. Named Entity Recognition and Resolution in Legal Text. In *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*. 27–43.

[36] Gary Edmond and Kristy A Martire. 2019. Just cognition: scientific research on bias and some implications for legal procedure and decision-making. *The modern law review* 82, 4 (2019), 633–664.

[37] Ahmed Elnaggar, Christoph Gebendorfer, Ingo Glaser, and Florian Matthes. 2018. Multi-Task Deep Learning for Legal Document Translation, Summarization and Multi-Label Classification. In *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*. 9–15.

[38] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.

[39] Frank Fagan. 2024. A View of How Language Models Will Transform Law. arXiv:2405.07826

[40] Atefeh Farzindar and Guy Lapalme. 2004. LetSum, an automatic Legal Text Summarizing system. In *Legal knowledge and information systems: JURIX 2004, the seventeenth annual conference*, Vol. 120. 11.

[41] Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal Judgment Prediction via Event Extraction with Constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[42] David Freeman Engstrom and Jonah B. Gelbach. 2021. Legal Tech, Civil Procedure, And The Future Of Adversarialism. *University of Pennsylvania Law Review* 169, 4 (2021), 1001–1099.
[43] Andrea Galassi, Francesca Lagioia, Agnieszka Jabłonowska, and Marco Lippi. 2024. Unfair clause detection in terms of service across multiple languages. *Artificial Intelligence and Law* (2024), 1–49.
[44] Andrea Galassi, Marco Lippi, and Paolo Torroni. 2023. Multi-Task Attentive Residual Networks for Argument Mining. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 31 (2023), 1877–1892.
[45] Debasis Ganguly, Jack G. Conrad, Kripabandhu Ghosh, Saptarshi Ghosh, Pawan Goyal, Paheli Bhattacharya, Shubham Kumar Nigam, and Shounak Paul. 2023. Legal IR and NLP: The History, Challenges, and State-of-the-Art. In *European Conference on Information Retrieval (ECIR)*. 331–340.
[46] Daphne Gelbart and JC Smith. 1991. Flexicon, a new legal information retrieval system. *Can. L. Libr.* 16 (1991).
[47] Dephne Gelbart and J. C. Smith. 1991. Beyond boolean search: FLEXICON, a legal tex-based intelligent system. In *Proceedings of the 3rd International Conference on Artificial Intelligence and Law*. 225–234.
[48] Joseph Gesnouin, Yannis Tannier, Christophe Gomes Da Silva, Hatim Tapory, Camille Brier, Hugo Simon, Raphael Rozenberg, Hermann Woehrel, Mehdi El Yakaabi, Thomas Binder, Guillaume Marie, Emilie Caron, Mathile Nogueira, Thomas Fontas, Laure Puydebois, Marie Theophile, Stephane Morandi, Mael Petit, David Creissac, Pauline Ennouchy, Elise Valetoux, Celine Visade, Severine Balloux, Emmanuel Cortes, Pierre-Etienne Devineau, Ulrich Tan, Esther Mac Namara, and Su Yang. 2024. LLaMandement: Large Language Models for Summarization of French Legislative Proposals. arXiv:2401.16182
[49] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023. DALE: Generative Data Augmentation for Low-Resource Legal NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
[50] John Gibbons and M. Teresa Turell. 2008. *Dimensions of Forensic Linguistics* (1 ed.). AILA Applied Linguistics Series, Vol. 5. John Benjamins Publishing Company. 1–317 pages.
[51] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview and Discussion of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. *The Review of Socionetwork Strategies* 18, 1 (2024), 27–47.
[52] Matthias Grabmair, Kevin D. Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R. Walker. 2015. Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. 69–78.
[53] S. Georgette Graham, Hamidreza Soltani, and Olufemi Isiaq. 2023. Natural language processing for legal document review: categorising deontic modalities in contracts. *Artificial Intelligence and Law* (2023).
[54] Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting Arguments in CJEU Decisions on Fiscal State Aid. In *Proceedings of the 9th Workshop on Argument Mining*. 143–157.
[55] Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law* 32, 3 (2024), 1–38.
[56] Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. In *Advances in Neural Information Processing Systems*, Vol. 35.
[57] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. arXiv:2103.06268
[58] Jia-Hong Huang, Chao-Chun Yang, Yixian Shen, Alessio M. Pacces, and Evangelos Kanoulas. 2024. Optimizing Numerical Estimation and Operational Efficiency in the Legal Domain through Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4554–4562.
[59] Weiyi Huang, Jiahao Jiang, Qiang Qu, and Min Yang. 2020. AILA: A Question Answering System in the Legal Domain. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.
[60] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2024. A sentence is known by the company it keeps: Improving Legal Document Summarization Using Deep Clustering. *Artificial Intelligence and Law* 32, 1 (2024), 165–200.
[61] Samyar Janatian, Hannes Westermann, Jinzhe Tan, Jaromir Savelka, and Karim Benyekhlef. 2023. From Text to Structure: Using Large Language Models to Support the Development of Legal Expert Systems. arXiv:2311.04911
[62] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. arXiv:2310.06825
[63] Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex 'Sandy' Pentland, Yoon Kim, Jad Kabbara, and Deb Roy. 2024. Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling. arXiv:2402.17019
[64] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. arXiv:1706.05137

[65] Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named Entity Recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 184–193.
[66] Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review* 51 (2019), 371–402.
[67] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 382, 2270 (2024).
[68] Soha Khazaeli, Janardhana Punuru, Chad Morris, Sanjay Sharma, Bert Staub, Michael Cole, Sunny Chiu-Webster, and Dhruv Sakalley. 2021. A Free Format Legal Question Answering System. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. 107–113.
[69] Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2024. A Survey on Challenges and Advances in Natural Language Processing with a Focus on Legal Informatics and Low-Resource Languages. *Electronics* 13, 3 (2024).
[70] Amirhossein Layegh, Amir H. Payberah, Ahmet Soylu, Dumitru Roman, and Mihhail Matskin. 2023. ContrastNER: Contrastive-based Prompt Tuning for Few-shot NER. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. 241–249.
[71] Jihoon Lee and Hyukjoon Lee. 2019. A Comparison Study on Legal Document Classification Using Deep Neural Networks. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. 926–928.
[72] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A Dataset of German Legal Documents for Named Entity Recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 4478–4485.
[73] LexisNexis [n. d.]. *International Legal Generative AI Report.* Retrieved July 22, 2024 from https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-international-legal-generative-ai-survey-shows-nearly-half-of-the-legal-profession-believe-generative-ai-will-transform-the-practice-of-law
[74] Jonathan Li, Rohan Bhambhoria, and Xiaodan Zhu. 2022. Parameter-Efficient Legal Domain Adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 119–129.
[75] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-Trained Language Models for Text Generation: A Survey. *ACM Comput. Surv.* 56, 9 (2024), 1–39.
[76] Yanling Li, Jiaye Wu, and Xudong Luo. 2024. BERT-CNN based evidence retrieval and aggregation for Chinese legal multi-choice question answering. *Neural Computing and Applications* 36, 11 (2024), 5909–5925.
[77] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law* 27 (2019), 117–139.
[78] Shuaiqi Liu, Jiannong Cao, Yicong Li, Ruosong Yang, and Zhiyuan Wen. 2024. Low-resource court judgment summarization for common law systems. *Information Processing and Management* 61, 5 (2024), 103796.
[79] Yang Liu. 2019. Fine-tune BERT for extractive summarization. arXiv:1903.10318
[80] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692
[81] Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. ML-LJP: Multi-Law Aware Legal Judgment Prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1023–1034.
[82] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[83] Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to Predict Charges for Criminal Cases with Legal Basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2727–2736.
[84] Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal Judgment Prediction with Multi-Stage Case Representation Learning in the Real Court Setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 993–1002.
[85] Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 130–142.
[86] Sepideh Mamooler, Rémi Lebret, Stephane Massonnet, and Karl Aberer. 2022. An Efficient Active Learning Pipeline for Legal Text Classification. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 345–358.
[87] Stelios Maroudas, Sotiris Legkas, Prodromos Malakasiotis, and Ilias Chalkidis. 2022. Legal-Tech Open Diaries: Lesson learned on how to develop and deploy light-weight models in the era of humongous Language Models. In *Proceedings of the Natural Legal Language Processing Workshop 2022*.
[88] Eric Martínez. 2024. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law* (2024), 1–24.
[89] Suzanne McGee. [n. d.]. *Generative AI and the Law.* Retrieved July 22, 2024 from https://www.lexisnexis.com/html/lexisnexis-generative-ai-story

[90] Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law* 28, 2 (2020), 237–266.
[91] Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. 1999. Abstracting of legal cases: the potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science* 50, 2 (1999), 151–161.
[92] Laurens Mommers. 2010. *Ontologies in the Legal Domain.* Springer Netherlands, 265–276.
[93] Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, and Paolo Italiani. 2023. Multi-language transfer learning for low-resource legal case summarization. *Artificial Intelligence and Law* (2023).
[94] Duy-Hung Nguyen, Bao-Sinh Nguyen, Nguyen Viet Dung Nghiem, Dung Tien Le, Mim Amina Khatun, Minh-Tien Nguyen, and Hung Le. 2021. Robust Deep Reinforcement Learning for Extractive Legal Summarization. In *Neural Information Processing*. 597–604.
[95] Ha-Thanh Nguyen, Manh-Kien Phi, Xuan-Bach Ngo, Vu Tran, Le-Minh Nguyen, and Minh-Phuong Tu. 2024. Attentive deep neural networks for legal document retrieval. *Artificial Intelligence and Law* 32, 1 (2024), 57–86.
[96] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*.
[97] Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 3016–3054.
[98] Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2024. MultiLegalPile: A 689GB Multilingual Legal Corpus. arXiv:2306.02069
[99] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
[100] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews* 10, 1 (2021).
[101] Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. 98–107.
[102] Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina Pantazi, and Manolis Koubarakis. 2021. Multi-granular Legal Topic Classification on Greek Legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. 63–75.
[103] Vedant Parikh, Upal Bhattacharya, Parth Mehta, Ayan Bandyopadhyay, Paheli Bhattacharya, Kripa Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2022. AILA 2021: Shared task on Artificial Intelligence for Legal Assistance. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*. 12–15.
[104] Sungmi Park and Joshua I. James. 2023. Lessons learned building a legal inference dataset. *Artificial Intelligence and Law* (2023).
[105] Sezen Perçin, Andrea Galassi, Francesca Lagioia, Federico Ruggeri, Piera Santin, Giovanni Sartor, and Paolo Torroni. 2022. Combining WordNet and Word Embeddings in Data Augmentation for Legal Texts. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 47–52.
[106] Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A System for Automated Summarization of Legal Texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. 258–262.
[107] Thiago Dal Pont, Federico Galli, Andrea Loreggia, Giuseppe Pisano, Riccardo Rovatti, and Giovanni Sartor. 2023. Legal Summarisation through LLMs: The PRODIGIT Project. arXiv:2308.04416
[108] Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal Corpus for Argument Mining. In *Proceedings of the 7th Workshop on Argument Mining*. 67–75.
[109] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text.. In *New Directions in Question Answering*. AAAI Press, 28–34.
[110] Vasile Păis, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. 9–18.
[111] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *The Review of Socionetwork Strategies* 16, 1 (2022), 111–133.
[112] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
[113] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108

[114] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108

[115] Piera Santin, Giulia Grundler, Andrea Galassi, Federico Galli, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2023. Argumentation Structure Prediction in CJEU Decisions on Fiscal State Aid. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*. 247–256.

[116] Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining Legal Concepts with Augmented Large Language Models (GPT-4). arXiv:2306.09525

[117] Marijn Schraagen, Floris Bex, Nick Van De Luijtgaarden, and Daniël Prijs. 2022. Abstractive Summarization of Dutch Court Verdicts Using Sequence-to-sequence Models. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 76–87.

[118] Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 31–46.

[119] Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. 2020. Large scale legal text classification using transformer models. arXiv:2010.12871

[120] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities. In *Advances in Neural Information Processing Systems*, Vol. 35. 13158–13173.

[121] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3407–3412.

[122] Juanming Shi, Qinglang Guo, Yong Liao, Yuxing Wang, Shijia Chen, and Shenglin Liang. 2024. Legal-LM: Knowledge Graph Enhanced Large Language Models for Law Consulting. In *Advanced Intelligent Computing Technology and Applications*.

[123] Răzvan-Alexandru Smădu, Ion-Robert Dinică, Andrei-Marius Avram, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2022. Legal Named Entity Recognition with Multi-Task Domain Adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*.

[124] Dezhao Song, Andrew Vold, Kanika Madan, and Frank Schilder. 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems* 106 (2022), 101718.

[125] Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. 2021. A dataset for evaluating legal question answering on private international law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 230–234.

[126] Francesco Sovrano, Monica Palmirani, Salvatore Sapienza, and Vittoria Pistone. 2024. DiscoLQA: zero-shot discourse-based legal question answering on European Legislation. *Artificial Intelligence and Law* (2024).

[127] Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. 2011. JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. 104–110.

[128] Zhongxiang Sun. 2023. A Short Survey of Viewing Large Language Models in Legal Aspect. arXiv:2303.09136

[129] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv:2102.02503

[130] Doron Teichman, Eyal Zamir, and Ilana Ritov. 2023. Biases in legal decision-making: Comparing prosecutors, defense attorneys, law students, and laypersons. *Journal of empirical legal studies* 20, 4 (2023), 852–894.

[131] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147.

[132] Suxin Tong, Jingling Yuan, Peiliang Zhang, and Lin Li. 2024. Legal Judgment Prediction via graph boosting with constraints. *Information Processing & Management* 61, 3 (2024).

[133] Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 1235–1241.

[134] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[135] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. arXiv:1710.10903

[136] Daniela Vianna, Edleno Silva de Moura, and Altigran Soares da Silva. 2023. A topic discovery approach for unsupervised organization of legal document collections. *Artificial Intelligence and Law* (2023).

[137] Qiqi Wang, Kaiqi Zhao, Robert Amor, Benjamin Liu, and Ruofan Wang. 2022. D2GCLF: Document-to-Graph Classifier for Legal Document Classification. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 2208–2221.

[138] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

[139] Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2022. Towards Interactivity and Interpretability: A Rationale-based Legal Judgment Prediction Framework. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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

[141] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. arXiv:1807.02478

[142] Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish Confusing Law Articles for Legal Judgment Prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3086–3095.

[143] Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 4085–4091.

[144] Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1854–1864.

[145] Ying Yin and Ivan Habernal. 2022. Privacy-Preserving Models for Legal Natural Language Processing. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 172–183.

[146] Mingruo Yuan, Ben Kao, Tien-Hsuan Wu, Michael M. K. Cheung, Henry W. H. Chan, Anne S. Y. Cheung, Felix W. H. Chan, and Yongxi Chen. 2023. Bringing legal knowledge to the public by constructing a legal question bank using large-scale pre-trained language model. *Artificial Intelligence and Law* (2023).

[147] Kwan Yuen Iu and Vanessa Man-Yi Wong. 2023. ChatGPT by OpenAI: The End of Litigation Lawyers.

[148] Gechuan Zhang, Paul Nulty, and David Lillis. 2023. Argument Mining with Graph Representation Learning. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*.

[149] Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. Contrastive Learning for Legal Judgment Prediction. *ACM Transactions on Information Systems* 41, 4 (2023).

[150] Weiqi Zhang, Hechuan Shen, Tianyi Lei, Qian Wang, Dezhong Peng, and Xu Wang. 2023. GLQA: A Generation-based Method for Legal Question Answering. In *2023 International Joint Conference on Neural Networks (IJCNN)*. 1–8.

[151] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 159–168.

[152] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal Judgment Prediction via Topological Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3540–3549.

[153] Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (2020), 1250–1257.

[154] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5218–5230.

[155] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A Legal-Domain Question Answering Dataset. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9701–9708.

[156] Yang Zhong and Diane Litman. 2022. Computing and Exploiting Document Structure to Improve Unsupervised Extractive Summarization of Legal Case Decisions. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. 322–337.

[157] Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson, and Felix Steffek. 2024. The cambridge law corpus: a dataset for legal AI research. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Received 24 September 2024; revised 29 July 2025; accepted 9 November 2025