IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 December 2024, accepted 19 January 2025, date of publication 23 January 2025, date of current version 29 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3533217

SURVEY

Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches

MARCO SIINO[1,2,3], MARIANA FALCO[2], DANIELE CROCE[2,3], AND PAOLO ROSSO[4,5]

[1]Department of Electrical, Electronics and Informatics Engineering, University of Catania, 95131 Catania, Italy
[2]Department of Engineering, University of Palermo, 90133 Palermo, Italy
[3]Palermo Research Unit, National Inter-University Consortium for Telecommunications (CNIT), 90128 Palermo, Italy
[4]PRHLT Research Center, Universitat Politècnica de València, 46022 Valencia, Spain
[5]ValgrAI–Valencian Graduate School and Research Network of Artificial Intelligence, 46022 Valencia, Spain

Corresponding author: Marco Siino (marco.siino@unict.it)

This work was supported in part by the European Union through the Italian Ministero dell'Università e della Ricerca (MUR) National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (program "Research and Innovation on Future Telecommunications Systems and Networks, to make Italy more Smart (RESTART)"), within the project Net4Future under Grant PE0000001; in part by the research project Malicious Actors Profiling and Detection in Online Social Networks through Artificial Intelligence (MARTINI) funded by Ministerio de Ciencia e Innovación (MCIN)/Agencia Estatal de Investigación (AEI)/10.13039/501100011033 under Grant PCI2022-135008-2; in part by the European Union NextGenerationEU/Plan de Recuperación, Transformación y Resiliencia (PRTR) and FairTransNLP-Stereotypes—Fairness and transparency for equitable NLP applications in social media: Identifying stereotypes and prejudices and developing equitable systems funded by MCIN/AEI/10.13039/501100011033 under Grant PID2021-124361OB-C31; and in part by the European Regional Development Fund (ERDF), EU, A way of making Europe.

ABSTRACT Artificial Intelligence (AI) is reshaping the legal landscape, with software tools now impacting various aspects of legal work. The intersection of Natural Language Processing (NLP) and law holds potential to transform how legal professionals, including lawyers and judges, operate, resolve disputes, and retrieve case information to formulate their decisions. To identify the current state of the applications of Transformers (also known as *Large Language Models* or *LLMs*) in the legal domain, we analysed the existing literature from 2017 to 2023 through a database search and snowballing method. From 61 selected publications, we identified key application categories such as legal document analysis, case prediction, and contract review, along with their main characteristics. We observed a discernible upsurge in the volume of scholarly publications, a diversification of tasks undertaken (e.g., legal research, contract analysis, and regulatory compliance), and an increased range of languages considered. There has been a notable enhancement in the methodological sophistication employed by researchers in practical applications. The performance of models grounded in the Generative Pre-trained Transformer (GPT) architecture has consistently improved across various legal domains, including contract review, legal document summarization, and case outcome prediction. This paper makes several significant contributions to the field. Firstly, it identifies emerging trends in the application of LLMs within the legal domain, highlighting the growing interest and investment in this area. Secondly, it pinpoints methodological gaps in current research, suggesting areas where further development and refinement are needed. Lastly, it discusses the broader implications of these advancements for real-world legal tasks, offering insights into how LLM-based AI can enhance legal practice while addressing the associated challenges.

INDEX TERMS Natural language processing, law, AI for law, legal NLP, legal tech, GPT, transformers, literature review.

I. INTRODUCTION

Law, as a discipline, continuously evolves in response to societal, political, economic, and technological changes [1]. The vast amount of information generated in the legal domain has prompted the exploration of machine assistance, leveraging computers' ability to analyse large textual datasets rapidly. Artificial Intelligence (AI), which refers to the simulation of human intelligence processes by machines, particularly computer systems, has become instrumental in this context.

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

AI encompasses capabilities such as learning, reasoning, and self-correction, enabling machines to perform tasks that typically require human intelligence. Today, AI is crucial for legal tech firms, enhancing operations to serve clients more affordably and accurately. Key AI applications include contract review, legal research, and predictive analytics for case outcomes and court rulings. Additionally, chatbots are becoming increasingly popular for providing self-service legal information. Political institutions and governments are actively supporting AI development. For instance, a 2020 report [2] highlighted AI's role in improving agency operations like regulatory enforcement and data analysis in the U.S. The European Union also emphasizes excellence and trust in AI, aiming to enhance research and industrial capacity while protecting fundamental rights [3].

The integration of Natural Language Processing (NLP) and AI into legal tasks is a natural progression, given the linguistic nature of law. This combination allows for more efficient and accurate analysis of legal texts, enhancing various aspects of legal practice. The use of NLP and AI in legal tech has a long history, dating back to the 1960s with the development of online legal content search systems [4], [5], [6]. NLP, particularly with the advent of Large Language Models (LLMs), has made significant strides in legal applications, aiding in tasks requiring language processing and understanding. LLMs represent cutting-edge technology, advancing AI approaches in various domains involving textual contents such as medicine [7], [8], [9], engineering [10], [11] and law [12], [13], [14], [15]. Overall, the integration of AI and NLP in the legal domain holds great promise for improving efficiency and decision-making processes across various legal tasks. However, addressing challenges related to context, data availability, and interpretability remains essential for the reliable application of these technologies in the legal domain [16], [17]. Within this context, applications refer to the broader use cases where AI technologies address specific legal challenges or fulfil specific needs, such as contract review, document automation, and compliance monitoring. These applications are representations of particular instances of how models, algorithms, or tools are applied to legal tasks. Conversely, tasks within legal AI are the specific activities that AI systems perform within these applications, such as identifying key clauses in contracts, extracting relevant information, or generating summaries.

This article aims to comprehensively examine the applications of LLMs in the legal domain, focusing on how various models and techniques are used in legal tasks. Based on recent literature [3], [17], [18], [19], [20], [21], [22], [23], [24], [25], we have grouped the tasks into three main areas: Legal Search, Legal Document Review and Legal Prediction. To identify the current state of the applications of Transformers in the legal domain, we analysed the existing literature from 2017 to 2023. The year 2017 was chosen as the starting point because it marks the introduction of the Transformer architecture [26]. A timeline of the evolution during these years is depicted in the Figure 1. We conducted a database search and snowballing method, resulting in an initial pool of publications. After a rigorous quality assessment and filtering process, we selected 61 publications that met our criteria for relevance and methodological rigour. By identifying and analysing specific instances where LLM-based methodologies are applied, we highlight their effectiveness and limitations in addressing various legal challenges. Our goal is to offer a thorough understanding of how these technologies can aid domain experts, such as law firms, judges, and lawyers, and facilitate automated resolution and document generation processes within the legal domain. This endeavour involves defining and extracting legal tasks and identifying trends and uses of LLMs in the legal domain.

The rest of the article is structured as follows: Section II presents the background and related work, and Section III describes the research method. Section IV reports a bibliometric analysis of the literature reviewed, Section V provides the overview of the models and approaches, whereas Section VI the trends within legal tasks and real-life implementations. Section VII states the conclusions, and Section VIII discusses the implications of our findings for future research.

II. BACKGROUND AND RELATED WORK

In this section, the landscape of legal AI is thoroughly examined in various studies, highlighting its diverse applications and advancements. The application of AI in the legal sector has long garnered significant interest, focusing on its implications for legal practice, administration, and the ethical considerations surrounding its use with legal data. Early explorations into online legal content search systems date back to the 1960s [5], [27], highlighting the longstanding intersection of AI and law. Zhong et al. [28] classify legal AI tasks into three categories: judgement prediction, similar case matching, and legal question answering, with early Transformer-based language models being notable contributors. Sansone and Sperlí [29] categorize legal information retrieval approaches into natural language-based, ontology-based, and deep learning-based systems. Additionally, Katz et al. [30] provide an extensive review of NLP in the legal domain, documenting the growth in research publications and tasks over the past decade. Dale et al. [31] focus on NLP applications in contract review and document automation, emphasizing their importance for legal practitioners.

The advent of Transformer models (from now on also LLMs) [26] marked a defining moment as well the LLMs, revolutionizing the landscape of deep-learning architectures and setting new benchmarks for performance across a spectrum of intricate NLP tasks. Central to the Transformer's innovation is its integration of attention mechanisms [32], which diverge from conventional Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). By harnessing attention mechanisms, the Transformer architecture orchestrates a holistic consideration of an

FIGURE 1. Timeline of key developments in Transformer models impacting the legal domain.

input context adeptly discerning the salient features while disregarding noise, thus enabling nuanced assessments of word dependencies regardless of their sequential proximity. The pervasive adoption of attention mechanisms underscores their pivotal role in NLP advancements, propelling significant improvements in task performance and scalability.

It represents a pioneering departure from traditional architectures by leveraging attention mechanisms, relying also on word embedding and neural networks. Attention is manifested in two distinct modalities within the Transformer framework. Initially, attention gauges the relative significance of words within the same sequence, treating input words as both origins and destinations. This self-attention mechanism endeavours to compute optimal representations of the input sequence, encapsulating syntactic and semantic relationships. Subsequently, input representations undergo weighted attention to predict target tokens in an autoregressive fashion.

LLMs such as ChatGPT demonstrate significant potential in various legal tasks, particularly in legal judgement prediction and statutory reasoning [33]. Trautmann et al. [34] introduce legal prompt engineering to enhance LLM performance in judgement prediction tasks, demonstrating effectiveness across multilingual datasets. Blair-Stanek et al. [35] explore GPT-3's aptitude for statutory reasoning, achieving high accuracy with dynamic few-shot prompting. Advancements in prompting techniques, like Chain-of-Thought (CoT) prompts presented by Yu et al. [36], further improve LLM performance in legal reasoning tasks. LLMs are also explored for their potential in legal education and supporting legal professionals [37]. Research by Iu Wong et al. [38] and Hargreaves [39] discusses the ethical utilization of AI in law school assessments, proposing methods to educate students on appropriate AI usage [40]. Pettinato [41] suggests that LLMs could assist law professors in administrative duties and streamline scholarly work. Macey-Dare [42] investigates LLMs as quasi-expert legal advisors, showcasing their feasibility in providing affordable legal counsel. In summary, LLMs have exhibited promising outcomes across diverse legal tasks, with advancements in prompting techniques playing a pivotal role in their efficacy. Nevertheless, challenges persist in ensuring the ethical utilization of LLMs and addressing their potential impact on the legal profession. Continued exploration of the capabilities and constraints of LLMs in the legal arena is essential, while ensuring their alignment with human values and societal requirements.

To systematically categorize the tasks in the legal domain where LLMs are applied, we adapted the structure proposed by Greco et al. [3], which divides legal challenges into three main areas: Legal Search [18], Legal Document Review [43] and Legal Prediction [17], [19], [20], [21], [22], [23], [24], [25].

Legal Search encompasses the following tasks:

- *Document Retrieval (T1)*: The process of finding relevant legal documents, such as case law, statutes, or legal opinions, from a large corpus of texts [17], [18], [44].
- *Case Entailment (T2)*: Retrieving documents that logically follows from or are supported by another case [18], [44].
- *Question Answering (T3)*: Answering specific legal questions by retrieving and synthesizing information from legal texts [18], [45].

Legal Document Review includes:

- *Named Entity Recognition (T4)*: Identifying and classifying entities (e.g., names of people, organizations, locations) within legal documents [43].
- *Similarity Estimation (T5)*: Measuring how similar two legal documents or cases are, which is essential for tasks like case law comparison [43], [46].
- *Classification (T6)*: Categorizing legal documents into predefined categories based on their content [24], [46], [47].
- *Document Summarization (T7)*: Producing concise summaries of lengthy legal documents [48], [49].
- *Datasets and Benchmarking (T8)*: Creating and using datasets to evaluate and compare the performance of various AI models in legal tasks [50].
- *Document Automation (T9)*: Automating the review and synthesis process of legal documents to enhance efficiency and accuracy [51].

Legal Prediction consists of:

- *Judgement Prediction (T10)*: Predicting the outcomes of legal cases based on previous rulings and case characteristics [19], [20], [21], [22], [52], [53].
- *Next Sentence Prediction (T11)*: Predicting the subsequent sentence in a legal document to assist in drafting and understanding legal texts [16], [17].

These tasks represent the specific activities that LLMs and other AI technologies perform within broader legal applications, facilitating advancements in legal research, document automation, and outcome prediction.

III. RESEARCH METHOD

In conducting our systematic review, we adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [54]. PRISMA is a widely accepted standard designed to help researchers conduct and report systematic reviews and meta-analyses in a transparent and comprehensive manner. It provides a checklist of items that should be included in the report of a systematic review or meta-analysis, ensuring that the review is conducted systematically and that all relevant information is clearly presented. By following the PRISMA guidelines, we aimed to enhance the quality and reliability of our review. Our methodology included a thorough search strategy, rigorous study selection criteria, systematic data collection, and comprehensive analysis. This approach ensured that our findings are robust and reproducible, providing a clear and transparent overview of the current state of LLMs in the legal domain. To identify relevant papers, we employed an adaptation of PRISMA guidelines, including a database search and complemented by the backward snowballing procedure [55], in order to reduce the risk of overlooking pertinent documents within the time frame selected. In the following subsections, we define the scope of the literature review through the research questions, the inclusion and exclusion criteria, the search and selection procedure, and the quality assessment criteria.

A. RESEARCH QUESTIONS

The scope of our literature review is related to the application of LLMs to the legal domain tailored to address both general and specialized tasks within the realm of law, and it is defined by the following Research Questions (RQs):

1) **RQ1**: *What are the NLP applications in the legal domain that have already made use of LLMs?*
2) **RQ2**: *Is the pre-training or the fine-tuning of LLMs specifically for the legal domain always beneficial? And in what cases and for what tasks?*
3) **RQ3**: *What are the main advantages and limitations of LLMs in the legal field?*
4) **RQ4**: *What are the possible applications of LLMs to the legal domain not yet fully explored?*

RQ1 serves as the overarching inquiry, aiming to identify the specific NLP tasks within the legal domain that have employed Transformer models. Building upon the initial question, RQ2 seeks to determine whether the pre-training of LLMs on general data, followed by fine-tuning them specifically for the legal domain, consistently yields beneficial outcomes in legal domains. Pre-training involves training the model on a large corpus of diverse texts to develop a broad understanding of language, while fine-tuning adjusts the model's parameters using a smaller, specialized dataset to improve its performance on specific tasks within the legal field. RQ2 also seeks to identify which of the NLP applications are currently viable for use in real-world legal settings. It involves assessing the maturity and performance of LLMs in these applications, examining case studies, pilot projects, and industry reports to identify instances where LLMs are successfully being used by legal professionals or institutions. Finally, RQ3 and RQ4 aim to explore advantages and limitations of LLMs in the legal domain, and legal tasks that have not yet been extensively researched or implemented. They aim to identify gaps in the current literature, brainstorming innovative uses of LLMs, and proposing new areas where these models could significantly impact legal practices, such as novel ways of automating legal processes, enhancing legal decision-making, or improving access to legal resources.

Overall, these research questions follow a logical progression, starting with broad inquiries about the usage of LLMs in the legal domain and gradually narrowing down to specific aspects and areas for improvement.

B. INCLUSION AND EXCLUSION CRITERIA

This subsection outlines the criteria used for assessing the relevance of the studies. We excluded all publications meeting any of the following criteria: (E1) studies unrelated to applications or tools of LLMs in the legal domain, (E2) non-English publications, (E3) non-research article types (e.g., type magazine, secondary study, course material, doctoral symposium paper, thesis, position paper, keynote presentation, abstract, or book chapter), (E4) inaccessible publications, and (E5) gray literature sources (e.g. blogs, manuals, tutorials, newsletter or project reports). Conversely, we included publications meeting at least one of the following criteria: (I1) studies addressing applications of LLMs in law, and (I2) relevant gray literature sources such as arXiv studies that fulfilled I1.

We evaluated the relevance of the studies by examining their titles, abstracts, and keywords. If the initial assessment was inconclusive, we delved deeper into the article's structure, introduction, methodology, results, and conclusions. The initial screening process was carried out by the first two authors. Any discrepancies or uncertainties were resolved through collaborative discussions among all four authors to achieve a consensus. It's important to note that while conducting the search, surveys, or systematic reviews emerged as results. Although these were assessed for relevance, they were included in the related section only if they aligned with our objectives, but they were not incorporated into the comprehensive list of analysed articles.

C. SEARCH AND SELECTION PROCEDURE

This subsection outlines the selection process. Regarding the database search, we selected a set of known repositories: IEEE Xplore, ACM Digital Library, SpringerLink, and Elsevier ScienceDirect, and we also complemented these with Google Scholar. Both searches were carried out during the last quarter of 2023. In order to refine the search, we formulated a combination of the following strings: *Legal AI*, *Legal NLP*, *NLP and Law*, *Case retrieval NLP* and *Contract review*, between 2017 to 2023.

It is pertinent to mention that the chosen timeframe aligns with the emergence of pivotal advancements in NLP, particularly marked by the introduction of the Transformer architecture [26] and the subsequent development of the BERT model [56], which occurred between the years 2017 and 2019.

We screened document titles and metadata to also apply the snowballing approach, and we removed duplicated records, records that were not accessible, that weren't applications or were written in a non-English language. As a result, we screened 307 studies and following the inclusion and exclusion criteria, we excluded 151 studies. 156 studies were assessed for eligibility and after the quality analysis we obtain the final set of 61 articles.

D. QUALITY ASSESSMENT CRITERIA

This section outlines the criteria utilized for assessing the quality of the included publications. We employed a checklist, as depicted in Table 1, constructed based on criteria outlined in the Critical Appraisal Skills Program (CASP)[1] and the methodology proposed in [57]. The questions in Table 1 are sequentially organized from Q1 to Q5. For each question, we devised sub-questions to facilitate the evaluation process for each publication. These sub-questions guided the assessment of publication quality, with responses categorized as either 'yes' or 'no', corresponding to values of 1 and 0, respectively. If any sub-question under a main question received a 'yes', the corresponding main question was also marked as 'yes'. Additionally, we deemed Q1 as the threshold criterion for further quality assessment, requiring a positive response to proceed.

TABLE 1. Questions of the quality assessment checklist.

ID	Topic	Description	Score
Q1	Reporting	Is there a clear statement of the goals and the application?	y/n
Q2	Rigour	Was the research design and methodology justified?	y/n
Q3	Rigour	Was the data collection and data analysis sufficiently rigorous?	y/n
Q4	Credibility	Does the study clearly state findings with credible results and justified conclusions?	y/n
Q5	Relevance	Does the study add value to existing research or practice?	y/n

IV. BIBLIOMETRIC ANALYSIS

After the quality assessment, we carefully selected a set of publications from which we were able to extract relevant data to answer the research questions. The resulting list of studies is composed of 61 publications, labelled as P1 to P61.[2] We use two types of references to ensure clarity and transparency. The [*ref number*] citations refer to the standard bibliographic entries listed at the end of the paper. Additionally, we use *Pxx* references to directly link to the publication details as collected during our research. These *Pxx* references correspond to the attached file containing all the selected papers, allowing readers to easily find and verify the specific publications discussed in our review. For example, P01 refers to the first publication in the attached file, while [1] refers to the first entry in the bibliography section. This dual referencing system aims to enhance the transparency and reproducibility of our work, providing readers with easy access to the relevant publication information. We believe this approach strikes a balance between streamlining the reading experience and maintaining the transparency and reproducibility of our research.

In this section we analyse the 61 selected publications in terms of venue type, geographic distribution, and author affiliation. These publications span from 2020 to 2023, showing a consistent upward trend in the amount of articles published each year regarding AI in the legal domain (10 studies in 2020, 15 in 2021, 17 in 2022, and 19 in 2023).

A. DISTRIBUTION BY TYPE OF VENUE

The majority of the selected publications were distributed across conferences (20 studies), workshops (13), and journals (7). There were also two articles presented in a single symposium (*JSAI International Symposium on Artificial Intelligence*) and one article in a forum (*Forum for Information Retrieval Evaluation*). Additionally, we identified 18 pre-prints available on arXiv. Among the conferences, the most prominent were the *International Conference on Artificial Intelligence and Law* (3 publications) and the *Conference on Neural Information Processing Systems* (3), followed by 14 different conferences. In terms of workshops, the *Natural Legal Language Processing Workshop* led with four publications, followed by the *Workshop on AI for Public Administration* with two, alongside seven other workshops. Notably, the dominant journal is *Artificial Intelligence and Law*, with four articles.

B. DISTRIBUTION BY AFFILIATION

We identified 37 studies within a purely academic context, 10 associated with various industries (encompassing companies, institutes, centers, and government entities), and 14 with shared affiliations. Among the 61 selected publications, authors are affiliated with 56 different universities, primarily located in Italy, Denmark, USA, India, Germany, and Greece. Additionally, affiliations include two government agencies (*Senato della Repubblica* and *Istituto Poligrafico e Zecca dello Stato*, Italy), 14 institutes, six centres (*Center for AI and Cognitive Computing at Thomson Reuters*, Canada; *Centre for AI*, University College London, UK; *German Research Center for Artificial Intelligence*; *Research Center for Digital Sustainability*, University of Bern; *CodeX*, Stanford Law School; and *iCourts*, University of Copenhagen), and one foundation (*Fondazione Bruno Kessler*, Trento,

[1]CASP Checklists, available at: https://casp-uk.net/casp-toolschecklists/ (Accessed: Dec. 28, 2024)

[2]The complete list of publications can be found at the following link: https://github.com/marco-siino/llm-applications-ieee-access (Accessed: Dec. 28, 2024)

Italy). Moreover, one project (The Atticus Project[3]) and 17 companies are represented. Notably, 14 studies had shared affiliations, which may involve partnerships between universities and companies, centres, institutes, projects, or foundations (P10, P11, P13-P15, P17, P19-P21, P23, P28, P29, P31, P32).

Additionally, we found two articles authored by multiple universities: the first one is authored by University of Bern, Bern University of Applied Sciences, University of Zurich, University of Bologna, University of Copenhagen, and Stanford University (P24), and the second one by ETH Zurich, Harvard University, The Atticus Project, The Nueva School, University of Wisconsin, Madison, Yale University, Stanford University, and UC Berkeley (P37). Furthermore, out of 56 universities, 27 appeared only once in the set of articles.

We also noticed a predominance of conferences published in a purely academic context, with 13 out of 20 articles. Likewise, workshops aligned with this trend, with seven out of 13 situated within academic settings. Regarding journals, four were exclusively affiliated with academia, while three had shared affiliations.

C. DISTRIBUTION BY COUNTRY

In terms of the geographic distribution of authors, Italy stands out with 15 publications, with three more featuring shared affiliations with Germany (P54), the UK (P61), as well as Switzerland, Denmark, and the USA (P24). Following Italy, the USA has seven publications, while India and the Netherlands each have three. Additionally, Canada, China, Denmark, Germany, Greece-UK, and Romania each have two publications to their credit. Moreover, five countries have only one article each: Brazil (P33), France (P56), Japan (P7), Korea (P3), and Singapore (P19).

Out of a total of 61 publications, 17 have shared affiliations. For instance, India collaborated with the USA (P5), the United Kingdom (P55), and Australia (P11). Switzerland also demonstrated shared affiliations with Germany (P14), Denmark (P2), the USA (P37), and the UK (P50). Furthermore, the United Kingdom has collaborated with Denmark (P42), Germany, the USA, Denmark, and Greece (P21), while the USA collaborated with Canada (P49). There are also instances of co-authorship between countries: Netherlands and Austria co-author one publication (P13), as do China and Japan (P28), and Greece and Denmark (P10).

It's worth noting the distribution of studies among universities and countries. For instance, in Denmark, we found only the University of Copenhagen with seven studies, followed by both University of Pisa, Italy and Stanford University, USA with five studies each. The distribution across Italy is also noteworthy, showcasing the broad range of regions engaged in this topic: University of Calabria (2), Scuola Superiore Sant'Anna, Pisa, Università degli Studi di Milano (2), Università per Stranieri di Siena (2), University of Trieste (1), Politecnico di Torino (1), Roma Tre University (1), Sapienza Università di Roma (1), University of Bari Aldo Moro (1), IUSS Pavia (1), and University of Bologna (1).

[3]The Atticus Project, Open-Source Dataset of Legal Contracts for AI. Available at: https://www.atticusprojectai.org/

D. DISTRIBUTION BY LEGAL DOMAINS

In our analysis, we discerned distinct legal domains addressed by each study, namely: case law, legislation, legal documents, legal research, and miscellaneous topics. These categories encompass various sub-areas grouped based on thematic similarities. Table 2 summarizes the areas, sub-areas and studies. It's feasible to mention that the Case Law area is mostly focused on court cases, including supreme court cases, European Court of Human Rights (ECHR) cases, civil cases, European Court of Justice cases, and supporting cases; showing interest in court decisions, opinions and precedents. In the Legislation area there is attention on national legislation, statutes, and civil law codes, with a significant focus on EU law, US law, and UK law, although also appeared the exploration of Indian Law. Then, in Legal Documents, there is an emphasis on contracts and case judgements, while there is also a notable interest in deal points, maxims and measures (e.g. COVID-19), and fewer studies showed interest on legal questions, legal news, and online legal resources.

TABLE 2. Distribution of legal areas, sub-areas and studies.

Area	Sub-areas	Studies
Case Law	Court cases	P02, P05, P06, P10, P11, P13-P15, P38, P48, P55, P60
	Supreme Court cases	P01, P24, P27, P40, P55
	ECHR cases	P21, P38, P52, P54, P56, P57
	Civil cases	P03, P25, P45, P59
	European Court of Justice cases	P38, P54
	Supporting cases	P28
	Court decisions	P01. P08, P24, P34, P54
	Court opinions	P22
	Court precedents	P03, P08, P09
	Negative outcomes	P50
	Holdings	P04, P58
Legislation	National legislation	P10, P36, P55, P57
	Statutes	P03, P06, P09, P11
	Civil law codes	P12, P16
	EU law	P21, P23, P34, P38, P57
	US law	P21, P33, P57
	UK law	P38, P55
	Indian law	P40, P55
	Public administration	P22, P29, P39, P43, P53, P60
	Legal acts	P30
Legal Documents	Company fillings	P42
	Contracts	P10, P17-P19, P21, P33, P35, P38, P57
	Case judgements	P20, P31, P40, P45, P46
	Redaction of legal documents	P20
	Online terms of service	P24, P56
Legal Research	Legal questions	P49
	Legal news	P31, P43
	Online legal resources	P40, P41, P43
Miscellaneous	Deal points	P37
	Legal Language	P41, P47
	Maxims	P31
	Measures	P24

V. MODELS AND APPROACHES

The LLMs application in law encompass a wide range of functionalities and objectives, such as contract review, legal research, document automation, case prediction, and compliance monitoring, among others. Furthermore, we define *application* as particular cases of how a model, an algorithm, a tool, or others is applied to a legal task.

To address the research questions, we approach legal AI tasks in two ways. First, we examine each legal task by reviewing the relevant studies, highlighting their key contributions, and identifying trends. Second, we summarize these tasks to offer valuable insights for the research community.

The rest of this section presents the main contributions to the identified legal tasks, with a summary of the tasks and corresponding studies shown in Table 3.

TABLE 3. Distribution of legal tasks and studies.

Area	Legal Tasks	Studies
Legal Search	Document Retrieval (T1)	P04-P10, P12-P16, P20, P28, P37
	Case Entailment (T2)	P06, P07, P13, P28
	Question Answering (T3)	P49, P53
Legal Document Review	Named Entity Recognition (T4)	P33-48
	Similarity Estimation (T5)	P04, P05, P09, P11, P12, P19, P20, P33
	Classification (T6)	P03, P10, P12, P21, P24, P26-P31, P38, P41, P44, P48, P50, P51, P52, P54, P55-P57
	Document Summarization (T7)	P03, P13, P27, P58, P59, P60
	Datasets and Benchmarking (T8)	P02-P04, P14, P17, P18, P21-P24, P32, P34, P35, P56, P58, P61
	Document Automation (T9)	P17-P19, P25, P35
Legal Prediction	Judgement Prediction (T10)	P01-P03, P32, P40, P50, P52
	Next Sentence Prediction (T11)	P39, P41, P42, P46

A. LEGAL SEARCH

1) DOCUMENT RETRIEVAL (T1)

In the legal domain, practitioners frequently require access to specific documents for tasks such as research, case preparation, drafting, and client advising. Document retrieval is pivotal in accessing relevant legal texts, including court cases, statutes, regulations, and legal opinions, to support these activities effectively [17], [18], [44].

Among the articles reviewed, 15 focused on Document Retrieval in the legal domain (P04-P10, P12-P16, P20, P28, P37), with a primary focus on Case Law and Legislation (see Table 2). In particular, 11 articles focus on Legal Case Retrieval, one on Statute Law Retrieval (P09), and two on both Legal Case Retrieval and Statute Law Retrieval (P06, P07). Notably, five articles specifically address long-document retrieval (P06, P07, P10, P13, P28).

Various tools, methods, frameworks, and language models have been proposed to enhance document retrieval. For instance, P05 [58] presents a Virtual Legal Assistant (VLA) that enables legal professionals to consult on legal situations with an AI-based assistant. The study emphasizes the Information Retrieval phase, refining search scope with increasing query intricacies and filters.

The COLIEE competition[4] stands out in this research field. In P06 [59], a tool with specialized multilingual search functionalities for document retrieval and entailment tasks was introduced. CatBoost was used for case law retrieval, and pre-trained embeddings combined with TF-IDF (i.e., Term Frequency–Inverse Document Frequency) [60] were employed for statutory information retrieval. The preprocessing included lowercase conversion, punctuation removal, and numeric digit-to-text conversion, with FastText for embedding training. Embedding-based methods proved to be more effective than those based on TF-IDF. A knowledge-based approach for legal document retrieval - relying on document embedding also in this case - and based on the organization of a textual data repository is introduced in [61]. Documents that have been pre-processed and embedded undergo iterative sentence-level classification through a cycle of terminology extraction and concept formation. This method leverages the ASKE (Automated System for Knowledge Extraction) engine, which tackles a multilabel classification problem without requiring any initial annotations of the documents. The authors present an application of ASKE in a practical case study focused on retrieving legal documents from a repository of Italian court decisions. This work is part of the Next Generation UPP (NGUPP) project, funded by the Italian Ministry of Justice, and aims to integrate AI and advanced information management techniques to facilitate the digital transformation of Italian legal processes and digital justice. Specifically, the case study addresses a common challenge faced by legal practitioners and administrators: the retrieval of past court decisions, known as ''precedents,'' based on one or more input text fragments, such as sentences, definitions, or excerpts from articles. The goal is to identify and retrieve the most relevant documents, such as court decisions or specific sentences within them, that match the input query. Similarly, P07 [62] addressed long documents and ambiguity in the legal domain, proposing a document-level attention mechanism and passage mining. They used abstract meaning representation to reduce noise and identified use-cases. The datasets comprised Canadian Federal Court cases and statute law. LEGAL-BERT was utilized to extract semantic relationships, and Spacy for text segmentation. TF-IDF outperformed BM25 in obtaining superior performance on long-document-related tasks. Regarding long-document tasks, in P10 [63], two methods for long-document processing were explored: modifying Longformer warm-started from LEGAL-BERT for longer texts (up to 8,192 tokens), and modifying

[4]Competition on Legal Information Extraction/Entailment (COLIEE), available at https://sites.ualberta.ca/ rabelo/COLIEE2024/

Legal-BERT to use TF-IDF representations. Experiments on LexGLUE showed these variants outperformed LEGAL-BERT, establishing new state-of-the-art solutions. Opposite than TF-IDF, using embedding representation P08 [64] introduced a knowledge-based approach for retrieving precedent sentences using document embedding models, with a zero-shot approach for classification and knowledge extraction from text fragments. Furthermore, in P20 [65] the authors introduce a novel method named PRILJ, designed to identify paragraph regularities in legal case judgments and assist legal experts in drafting legal documents. PRILJ employs a two-step approach: first, it groups documents into clusters based on their semantic content, and then it identifies regularities within the paragraphs of each cluster. The method utilizes embedding techniques to represent documents and paragraphs in a semantic numerical feature space. Additionally, an approximated nearest neighbor search method is used to efficiently retrieve the most similar paragraphs relative to those in a document being prepared. Another successful application of embedding representation is reported in P13 [66]. It combined lexical and dense retrieval methods for paragraph-level case retrieval. The best results were achieved with BM25 and dense passage retrieval using domain-specific embeddings. BM25 was used for initial retrieval, followed by BERT for aggregation and reordering. A BERT variation is discussed in P28 [67] where the authors proposed BERT-PLI to model paragraph-level interactions and infer relevance between cases, using a cascade framework to reduce computational costs. The model was fine-tuned with a small-scale case law entailment dataset, demonstrating effectiveness in legal scenarios. Also based upon BERT, P09 [68] focused on information retrieval systems to identify relevant precedents and statutes using text similarity approaches.

Domain-specific adaptation of BERT are often accomplished pre-training or fine-tuning the original BERT model on specific domains. For instance, P12 [69] introduced LamBERTa, a BERT-derived model pretrained on Italian Civil Code laws, highlighting predictive justice. LamBERTa predicts relevant articles from the Italian Civil Code in few-shot scenarios, emphasizing domain-specific adaptation. P16 [70] revised LamBERTa, incorporating legal-specific pre-training and out-of-vocabulary legal terms. Extensive evaluation revealed significant improvements in law article retrieval due to domain- and task-adaptation.

Some variations of the original Transformer architecture have also been found in the literature. For example, P15 [71] introduced SAILER, a structure-aware pre-trained language model for legal case retrieval, integrating structural information and an asymmetric encoder-decoder architecture. SAILER outperformed previous methods in legal case retrieval tasks, demonstrating strong discriminative capabilities.

Finally, in P14 [72], the authors evaluate 27 methods (based on both Transformers and traditional architectures) on standard benchmarks derived from Open Case Book and Wiki source. The objective of the task is to retrieve relevant literature given a specific case. The methods assessed in the study can be broadly categorized into three groups: word-vector based, Transformer based and citations based.

What emerged from the review of the literature is that performing fine-tuning or pre-training on a Transformer-based architecture as BERT, or on a word embedding representations, is often beneficial and common in the field of article and case retrieval. Some exceptions are traditional approaches, as TF-IDF. However, we barely found foundational models able to outperform domain-specific models and approaches for retrieval tasks.

2) CASE ENTAILMENT (T2)

Case Entailment is the task of determining whether the facts, legal principles and arguments introduced in a legal case logically support or imply the outcome of another case, requiring to analyse the content of legal documents to identify relationships and inferring conclusions based on precedents [18], [44]. We have found 4 studies that focus on this task (P06, P07, P13, P28), and their main legal area of the application is Case Law.

In P06 [59], out of 4 tasks, Task 2 (a case law entailment task) aims to distinguish which paragraph in a supporting case implies the provided text fragment and Task 4 (statutory entailment task) focuses on determining the potential implication of a bar exam question by a set of relevant articles. For Task 2, the jurisprudence implication task requires finding an implying paragraph from a case, given a base case with a specified fragment f. Three groups of features (traditional, embedding similarity, and Natural Language Inference (NLI)) were used also in this case. An ensemble classifier (i.e., XGBoost) was employed, and for each related case R, results were ranked based on the number of paragraphs implying f. The best embedding results were obtained with BERT. With this model and on this task, the authors achieved satisfactory results in the competition. The Task 4, the statutory law implication task, on the other hand, involves determining whether a legal bar exam question Q is included in the text of a series of articles $S1, S2, \ldots SN$ relevant to Q. Implication means that Q is true or false based on the content of $S1, S2, \ldots SN$. This objective was pursued in two ways: using a BERT-XGBoost combination and employing legal embeddings with a Bi-GRU. An interesting consideration is the results achieved by operating directly on the Japanese language, which typically contains more informative tokens than their English counterparts [73], [74]. Their findings illustrate that using legal embeddings and auxiliary linguistic features, such as NLI, showed the most promise for future improvements.

Also in P13 [66] the authors took advantage of a domain-focused embedding. The authors detail their methodologies for two specific tasks in the COLIEE 2021 competition: legal case retrieval and legal case entailment. The objective of the

legal case entailment task is to develop a system capable of identifying paragraphs within a relevant case that support the decision of a new, given case. In this task, a query paragraph is provided along with candidate paragraphs from a legal case, and the system must pinpoint which candidate paragraphs logically support the decision outlined in the query paragraph. The datasets used for training and testing include cases from the Federal Court of Canada. To identify the entailing paragraphs p for a given query paragraph q, the authors employ both lexical and semantic ranking techniques to rank the candidate paragraphs. They evaluate the performance of two approaches: BM25 and lawDPR. The most effective results were achieved by combining BM25 with dense passage retrieval, utilizing domain-specific embeddings to enhance accuracy.

It is also worth reporting that for this task (i.e., *entailment*) the most recurrent approaches are based on domain-specific embedding or BERT adaptation to the legal domain. Even if the tasks related to entailment are not yet fully solved in the literature, LLM-based approaches have consistently outperformed traditional and statistical method. This is due to their ability to identify contextual and semantic relations between words.

3) QUESTION ANSWERING (T3)

This task is focused on the application of NLP and machine learning techniques to automatically provide answers to legal queries, which involves understanding and interpreting complex legal questions and retrieving or generating accurate and relevant answers based on legal texts such as statutes, case law, regulations, and legal opinions [18], [45]. We have found two studies for this task (P49 and P53) which are embedded within the Legal Research and Miscellaneous legal areas.

Understanding legal texts presents a considerable challenge due to their extensive and intricate clauses, compounded by a scarcity of datasets annotated by experts. To tackle this issue, P37 [75] introduced the Merger Agreement Understanding Dataset (MAUD), a reading comprehension dataset meticulously annotated by experts based on the American Bar Association's 2021 Public Target Deal Points Study.[5] With over 39,000 examples and more than 47,000 annotations, MAUD serves as a significant resource. The authors assert that MAUD stands as the sole expert-annotated merger agreement dataset, making it an invaluable benchmark for both the legal profession and the NLP community. Their fine-tuned LLMs displayed promising performance, consistently outperforming random chance on most questions.

In P49 [76], a question-answering system is introduced to address legal queries. Sparse vectors and embeddings are employed as input for a BERT-based answer ranking model. The distinction is made between factoid and non-factoid questions. The proposed model is designed to address legal questions across a wide range of legal domains. The system selects and ranks answers using a combination of sparse vector techniques like BM25 and dense vectors (semantic embeddings). The answer selection process is conducted on a collection comprising over 100 million passages. The system employs Query By Document (QBD) implemented with BM25 for sparse vector representation, and Legal GloVe and Legal Siamese BERT for semantic embeddings. The values on the answer evaluation scale serve as a threshold to determine which answers to accept and which to reject. Many of the solutions implemented in this study hold notable relevance and are of interest for future developments.

In P53 [77], the authors introduce FRAQUE, a system designed to answer factual questions within the Public Administration sector. This system utilizes semantic frames, which are structured collections of slots with defined possible values. FRAQUE operates by querying unstructured data from various sources such as documents, websites, and social media. Leveraging statistical components like word embeddings, the system offers flexibility in adapting to different domains and languages. The primary objective of FRAQUE is to match questions with relevant frames and corresponding document passages stored in a knowledge graph, which are then presented as answers. To ensure user-friendliness, FRAQUE's development follows a user-centered design approach, allowing for the monitoring of linguistic patterns used by users and identifying the most frequently occurring structures in their queries.

Both P49 [76] and P53 [77] highlight the importance of developing question-answering systems for legal and public administration domains. The use of sparse vector techniques like BM25 and dense vectors (semantic embeddings) in P49 and pattern-based systems in P53, have shown promising results in addressing legal and factoid questions, respectively. The user-centred design process employed in P53 ensures the system's usability and adaptability to different domains and languages. Overall, the development of such systems has the potential to significantly improve the efficiency and accuracy of legal and public administration processes.

B. LEGAL DOCUMENT REVIEW

1) NAMED ENTITY RECOGNITION (T4)

Named Entity Recognition (NER) is a crucial stage in many legal AI tasks, facilitating the organization of unstructured text data into a more structured and searchable format [43]. We reviewed 16 studies that contribute to the advancement of NER in the legal domain. These studies encompass the introduction of new datasets, development of NER systems, creation of language models, evaluation of models, and proposition of legal annotation procedures. The primary focus of these studies is Case Law, although other legal areas are also covered.

Several studies have introduced new datasets for NER. These include a dataset comprised of German federal court

[5]https://www.americanbar.org/groups/business_law/about/committees/mergers-and-acquisitions/deal-points/ (Accessed: Dec. 28, 2024)

decisions (P34 [78]), which were evaluated using Conditional Random Fields (CRF) and Bidirectional Long Short-Term Memory (Bi-LSTM). Another dataset, CUAD, was curated for legal contract review (P35 [79]), with experiments conducted using Transformer-based architectures such as BERT, RoBERTa, AlBERT, and DeBERTa. Additionally, MAUD, a dataset based on the 2021 American bar examination, was presented in (P37) [75]. P42 utilized a publicly available legal NER dataset (ENER), with a subset of filings extracted for training different NER algorithms on the general English CoNLL-2003 corpus.

In terms of NER system development, P36 [80] presented a system within the Romanian legal domain. The authors curated a manually annotated corpus, MARCELL-RO, which encompasses legal documents extracted from the extensive *MARCELL project*.

Regarding language models, P38 [81] introduced LEGAL-BERT, a family of language models adapted to the specific legal domain through domain-specific pre-training or trained from scratch. The performances achieved were superior to using BERT-base. The same study introduced LEGAL-BERT-SMALL, with experiments conducted on several datasets, including one for NER focused on USA contracts.

P39 [82] extended the base vocabulary of the Italian Transformer model UmBERTo, resulting in BureauBERTo, through further pre-training on a corpus of documents related to the Italian Public Administration, banking, and insurance sectors. Two evaluation strategies were employed: an intrinsic strategy involving predicting masked words within the sequence, and an extrinsic approach involving a specific NER task tailored to the Public Administration domain. Also related to the Italian domain, in P41 [83], the authors evaluated NER tasks using ITALIAN BERT + Spacy NER and ITALIAN-LEGAL-BERT + Spacy NER. For the semantic similarity task, a subset of sentences from the Italian Civil Law DB was utilized.

A novel architectural approach based on a hierarchical Bi-LSTM was proposed in P40 [84], with the SemEval task encompassing three subtasks: RR Prediction, L-NER, and CJPE. The proposed model, HLBERT-CRF, incorporates both word-level and sentence-level encoders. For tokenization, a LEGAL-BERT-compatible tokenizer was employed, and F1-score and weighted F1 were used as evaluation metrics.

P42 [85] tested various NER algorithms on the CONLL-2001 corpus and evaluated them using the provided dataset collection. The paper also describes the E-NER corpus, an annotated collection of legal documents. It contains detailed annotations for entities like case numbers, court names, statutes, and legal parties, making it a valuable resource for training models that can accurately extract critical information from complex legal texts, streamlining tasks such as case law research and contract analysis. In P44 [86], a comparison was made among five Transformers pretrained on general-purpose text for two tasks within the domain of Public Administration in Italy: NER and Multi-Label Document Classification. The experimental results revealed that UmBERTo outperformed other Transformers for both tasks. The Transformers considered were BERT-BASE-ITA, UmBERTo, Multilingual BERT, XLM-RoBERTa, and GePpeTto, the latter being the first autoregressive Italian language model built upon GPT-2. Furthermore, the authors introduced a novel dataset specifically for this task, derived from a corpus they constructed, named the ''ATTO'' Corpus. The results consistently demonstrated UmBERTo's superior performance across various tasks and dataset variants.

In conclusion, the recent studies reviewed in this section have provided significant contributions to the advancement of NER in the legal domain. The introduction of new datasets, such as those comprised of German federal court decisions, CUAD, MAUD, and E-NER, has expanded the resources available for training and evaluating NER models. The development of NER systems, like the one presented in P36 [80], has improved the accuracy and efficiency of entity recognition within legal documents. The creation of language models, such as LEGAL-BERT and BureauBERTo, has further enhanced the performance of NER tasks by adapting to the specific legal domain and extending base vocabularies. The evaluation of these models has demonstrated their superiority over general-purpose language models in handling legal text. Moreover, the proposition of novel architectural approaches, like the hierarchical Bi-LSTM in P40 [84], has opened new avenues for improving the structural understanding of legal documents. The comparison of various Transformer models in P44 [86] has also highlighted the effectiveness of domain-specific models, such as UmBERTo, in handling legal NER tasks.

However, despite these advancements, challenges remain. The complexity and variability of legal language, as well as the need for large annotated corpora, continue to pose obstacles. Future research should focus on addressing these challenges, further improving the performance and applicability of NER in the legal domain.

2) SIMILARITY ESTIMATION (T5)

In the context of legal AI, this tasks focus on measuring how closely related or alike two pieces of legal text are, which can be comprised of comparisons between various types of legal documents such as court cases, statutes, contracts and legal opinions [43], [46]. We have identified 8 studies that approach this task (P04, P05, P09, P11, P12, P19, P20, P33), and most of the studies are embedded within the Case Law and Legislation areas.

One common theme among these studies is the importance of domain-specific pretraining and the use of text similarity techniques. In P04 [23], the authors introduced a dataset called CaseHOLD and evaluated the performance of different models on it. The results showed that domain pretraining with a custom legal vocabulary exhibited the most substantial performance gains, with a 7.2% gain on F1, representing a 12% improvement over BERT. This approach also demonstrated consistent performance gains across other

legal tasks. Similarly, in P09 [68], the authors proposed a text similarity approach for retrieving previous cases and statutes. They applied three variations of word representation based on Glove, Doc2Vec, and TF-IDF methods. The experiments demonstrated that the TFI-DF method achieved reasonable results compared to Doc2Vec and Glove methods, which usually require large training datasets. In P11 [87], the authors tackled the challenge of measuring the similarity between two legal cases. They introduced Hier-SPCNet, an enhancement of PCNet that incorporates a heterogeneous network of statutes or written laws relevant to the jurisdiction. The experiments were conducted using data from the Indian judiciary, where the benchmark similarity between document pairs was assessed by legal experts from two esteemed law institutes in India. The findings demonstrated that Hier-SPCNet achieved state-of-the-art performance in network-based legal document similarity.

Another common point among these studies is the use of document clustering and topic modelling techniques. In P12 [69], the authors created LamBERTa, a specialized deep learning framework for civil-law codes, specifically trained using the Italian civil code. This framework involves refining an Italian pre-trained BERT model on either the entire Italian civil code or parts of it, treating the retrieval of law articles as a classification task. The authors utilized a centroid-based partition clustering algorithm on a document-term matrix. They constructed a vectorial Bag-of-Words (BoW) model within the term feature space, employing TF-IDF for term relevance weighting and cosine similarity to compare documents. In P33 [88], the authors evaluated the use of BERTopic for topic modelling in legal documents. The researchers concentrated on a selection of landmark cases from the US Caselaw dataset to assess the impact of topic modelling, utilizing domain-specific embeddings pre-trained with LEGAL-BERT. Their findings indicate that incorporating references to statutory law, such as the US Code, during the text embedding process enhances the quality of topic modelling.

In conclusion, these studies highlight the potential of text similarity, document clustering, and topic modelling techniques for legal document analysis. The proposed methods have shown promising results in improving the performance of legal document analysis tasks, such as law article retrieval, case retrieval, and topic modelling. However, further research is needed to address the challenges posed by the complexity and variability of legal language and the need for large annotated corpora. The development of more sophisticated techniques for legal document analysis could have significant implications for the legal industry, enabling more efficient and accurate analysis and retrieval of legal documents.

3) CLASSIFICATION (T6)

Through the classification task, it is possible to categorize legal texts or documents into predefined classes or categories based on their content, using machine learning and NLP techniques to automate the organization and analysis of large volumes of legal information [24], [46], [47]. We have found 22 studies (P03, P10, P12, P21, P24, P26-P31, P38, P41, P44, P48, P50, P51, P52, P54, P55-P57) that refer to this task, where Case law is the most frequently addressed area, appearing in 16 articles, indicating its critical importance in legal research. Legislation is also a significant focus, present in 13 articles, reflecting the ongoing need to understand and interpret statutory laws. Legal documents are prominently featured in 10 articles, emphasizing the necessity of understanding various legal texts. Additionally, miscellaneous topics and legal research are covered in a few articles, highlighting emerging areas and interdisciplinary approaches within legal studies.

Several studies have proposed different approaches to tackle various legal text classification tasks. Since it is not feasible to discuss all the existent literature, in this paragraph we provide an overview of four representative studies, namely P03, P12, P24, P26 and P27 and highlights their findings.

P03 [89] introduced LBOX OPEN, a dataset consisting of a corpus and two classification tasks, namely case name and statute prediction from the factual description of individual cases. The authors also released realistic variants of the datasets by extending the domain to infrequent case categories in case name and statute classification tasks.

P12 [69] presented LamBERTa, a model derived from BERT and pre-trained on Italian Civil Code laws. Through a classification stage, the model predicts the most relevant articles from the Italian Civil Code. The study evaluated two learning approaches, a global one based on the entire Italian Civil Code, and a local approach based on individual books. The results showed that the local learning-based method outperformed the others, and LamBERTa exhibited generally superior performance compared to the other models considered.

P24 [90] introduced LEXTREME, a new benchmark for evaluating NLP models on legal tasks. The benchmark comprises three selections of 11 datasets spanning a total of 24 languages. The three classification tasks are Single Label Text Classification (SLTC), Multi Label Text Classification (MLTC), and Named Entity Recognition (NER). The best-performing baseline model (XLM-R) achieved an aggregated final score of 61.3.

P26 [47] applied the sliding window concept based on attention, previously proposed with Longformer, to DistilBERT. The proposed model consists of 6 Longformer attention heads and a total of 69 million parameters. The maximum context window size is increased to 4096, allowing the model to process texts up to 8 times longer than standard BERT [91], [92]. The authors used a pre-training dataset of 8GB, consisting of five types of documents, and applied two pre-training strategies. The results showed that the proposed model outperformed BERT, RoBERTa, and DistilBERT.

Finally, P27 [93] addressed the classification of legal cases by employing BERT, RoBERTa, Legal-BERT, Longformer, and LegalFormer. The authors divided the documents into segments with a maximum size of 512 tokens and selected the text fragment that yielded the best result given the considered metric. The results showed that Legal-BERT and Legal-Longformer achieved the best performance.

In conclusion, the studies discussed in this paragraph propose different approaches to tackle various legal text classification tasks. The findings suggest that pre-training models on domain-specific data and employing attention-based mechanisms can significantly improve the performance of legal text classification. However, it is worth noting that some studies lack comparisons with more contemporary iterations of GPT models, and further research is needed to explore their potential in legal tasks.

4) DOCUMENT SUMMARIZATION (T7)

This task allows generating concise and coherent summaries of lengthy legal documents, extracting the most relevant information from legal texts, such as court cases, statutes and contracts, making it easier to understand for legal professionals [48], [49]. We identified 6 studies (P03, P13, P27, P58, P59, P60) that have considered this task, and all of them are focused on *Case Law* as legal area.

A recent study - P03 [89] - presented LBOX OPEN, a legal corpus with Korean precedents, along with a summarization task consisting of Supreme Court precedents and their corresponding summaries. Another study - P58 [83] - introduced a system for extracting legal holdings from Italian judgments. The problem was modelled as a summarization task, with a focus on identifying salient passages. The authors curated a new dataset, ITA-CaseHold, containing over 1,100 pairs of judgments and their corresponding holdings. The authors fine-tuned Italian-LEGAL-BERT-SC to predict the most relevant sentences within a document, and the proposed model outperformed other baseline models in terms of Rouge R-1 and R-2 metrics.

P59 [94] addressed the Civil Rights Litigation Clearinghouse (CRLC) and introduced Multi-LexSum, a collection of 9,280 summaries authored by domain experts and derived from CRLC documents. The authors demonstrated that state-of-the-art models performed inadequately on this dataset. The authors employed the best-performing Transformers for summary generation, including BART, PEGASUS, PRIMERA, and LED. However, human evaluation of the generated summaries revealed an average modification of 87 tokens per paragraph, 76% paragraph length modification, and 65% total length modification, highlighting the unsatisfactory performance of all the Transformers considered in comparison to expert-authored summaries.

To address the challenge of limited data annotated by experts, P60 [95] introduced extractive summarization techniques for legal decisions in a low-resource environment. The dataset consisted of 112 decisions and their corresponding expert-annotated summaries. The authors leveraged Legal-BERT, pre-trained on Harvard's legal case corpus, and proposed a multitask model to distinguish Reasoning/Evidence from other rhetorical roles. Comparative baselines included MMR and TextRank, and evaluation metrics encompassed ROUGE-1 and ROUGE-2.

In conclusion, the studies discussed in this paragraph highlight the potential of NLP techniques for summarizing legal documents. However, the results also indicate that there is still room for improvement in terms of summary quality and accuracy, particularly in low-resource environments and for complex legal documents. The use of domain-specific pre-trained models, such as LEGAL-BERT, and the creation of new datasets, such as ITA-CaseHold and Multi-LexSum, are important steps towards improving the performance of NLP techniques in summarizing legal documents.

5) DATASETS AND BENCHMARKING (T8)

Datasets and Benchmarking help to measure and compare the performance of AI models in the legal domain, and also foster the creation of new datasets that can become available for the community; these are crucial elements for advancing research, development, and application of AI technologies in legal contexts [50]. We have found 15 studies that focused on this field, most of them introducing new datasets (P02, P03, P04, P17, P21, P22, P23, P32, P34, P35), proposing new benchmarks (P24, P56,), new benchmark datasets (P14, P18, P58) or new corpus (P61). Table 4) summarized the datasets introduced that can be categorized as court cases (P02, P14, P32), court precedents (P03), and various legal documents such as holdings (P04), contracts (P17, P22, P35), and facts from ECHR rulings (P21, P56).

Additionally, there is a strong emphasis on developing benchmarks for court precedents, case names, statutes, and various types of legal decisions (P03, P24, P58). Contract review is another area of interest, with studies focusing on analysing contract terms and conditions (P18, P35). Furthermore, NER datasets are being created to improve the extraction of information from court decisions (P34 [78]). The trend also includes benchmarking legal texts using support vector machine classifiers on the LexGLUE dataset, which encompasses facts from ECtHR rulings, court opinions, and online terms of service (P56 [96]). Overall, these trends underscore the critical role of creating and evaluating new datasets and benchmarks to advance the capabilities of legal AI in handling diverse legal documents. The performance and effectiveness of LLMs in legal domain are evaluated using a combination of standardized benchmarks and custom criteria tailored for the legal domain. Standardized benchmarks such as LexGLUE and CUAD provide a robust framework for comparing the performance of different LLMs across various legal NLP tasks. LexGLUE includes tasks like legal judgment prediction, statute law retrieval, and contract review, while CUAD is specifically designed for contract review and understanding. In addition

to these benchmarks, custom criteria that are tailored to the specific needs of the legal domain are present in the literature. These include domain-specific metrics such as precision, recall, and F1-score for tasks like named entity recognition (NER) in legal documents, legal text classification, and case outcome prediction. Also, evaluations by legal experts who assess the practical relevance and accuracy of the outputs generated by LLMs are reported in the literature. This expert evaluation helps ensure that the models meet the high standards required in legal applications. Furthermore, given the importance of context in legal texts, the models' ability to understand and generate contextually accurate responses is also evaluated, which is particularly crucial for tasks like legal document summarization and contract analysis. By using a combination of standardized benchmarks and custom criteria, the aim is to provide a comprehensive assessment of the performance and effectiveness of LLMs in legal applications.

One of the pivotal datasets within the legal domain is introduced in P21 [50] which represents a compilation of distinct sub-datasets aimed at assessing the performance across various Natural Language Understanding (NLU) tasks within the legal domain. It also offers an evaluation of several generic and legal-oriented models. The LexGLUE dataset comprises seven NLP datasets that are domain-specific, thoughtfully selected using SuperGLUE criteria. The most prevalent task pertains to the prediction of legal case outcomes. Subsequent tasks include Topic Classification, Information Extraction, Legal Question/Answering, and others. In the initial release of this dataset, the sole language considered was English. Beyond the utilization of Transformers, this study also encompasses the evaluation of Support Vector Machines (SVM). The SVM model's performance was fine-tuned using TF-IDF representation and hyperparameter optimization through grid search. Notably, the SVM exhibited particularly promising results, although LEGAL-BERT generally outperformed other models across all datasets. The LexGLUE dataset and benchmark are readily accessible via Hugging Face, along with the associated experimental code.[6]

In P22 [97], the authors present an expanding dataset, currently encompassing approximately 256 gigabytes of data. This dataset comprises a wide array of legal and administrative documents, including judgements, contracts, and legal materials released by the European Parliament. Notably, they initiated their work from a BERT-based model. A primary objective of this endeavour is the creation of a dataset devoid of privacy violations and toxic content. This ensures that any model, be it a LLM or a traditional one, can be trained free from biases. The presented results have shown that smaller models pretrained on domain-specific data exhibit superior performance compared to their more broadly pretrained, larger counterparts.

[6]LexGLUE dataset on Hugging Face, available at: https://huggingface.co/datasets/lex_glue (Accessed: Dec. 28, 2024)

TABLE 4. Datasets and benchmarking: Legal areas and contributions.

Study	Type	Name	Component	Language
P02	Dataset		Court cases	German, French, Italian
P03	Benchmark	LBOX OPEN	precedents, case names, statutes, civil cases	Korean
P04	Dataset	CaseHOLD	Qs to identify holdings	English
P14	Benchmark dataset			
P17	Dataset		Contracts	English
P18	Benchmark dataset		lease agreement	
P21	Datasets		facts from ECtHR rulings, court opinions, EU law documents, terms of services	English
P22	Dataset		court opinions, contracts, administrative rules, and legislative records	English
P23	Dataset		EU laws	23 languages
P24	Benchmark	LEXTREME	Court decisions, legal code, Supreme Court Cases, online terms of service, measures, EU laws	24 languages
P32	Dataset		Court cases	Romanian
P34	Dataset		Court decisions	Germany
P35	Dataset	CUAD	Contracts	English
P37	Dataset	MAUD	Merger agreements	English
P56	Benchmark		facts from ECtHR rulings, court opinions, EU law documents, terms of services	English
P58	Benchmark dataset	ITA - CaseHOLD		English, Italian
P61	Corpus		administrative language	Italian

In P23 [98], a corpus of 65,000 laws pertaining to the European Union is employed. Taxonomic labels are derived from Eurovoc.[7] These laws, issued by the single European countries, undergo translation from the 23 official languages. A significant portion of the dataset pertains to the issue of contractual advantage imbalance between contracting parties. The dataset is meticulously stratified, comprising over 50,000 samples in the training set spanning the period from 1958 to 2019. Additionally, approximately 5,000 samples are allocated both to the development set, covering the years 2010 to 2012, and the test set, from 2012 to 2020. Notably, this dataset is readily available on Hugging Face for seamless integration with Transformer models.[8]

In P14 [72], the authors evaluate 27 methods, including both Transformer-based and traditional architectures, using

[7]Eurovoc, available at: https://eur-lex.europa.eu/browse/eurovoc.html (Accessed: Dec. 28, 2024)

[8]Eurovoc on Hugging Face, available at: https://huggingface.co/DSs/multi_eurlex (Accessed: Dec. 28, 2024)

standard benchmarks derived from Open Case Book and Wiki source. The task's objective is to retrieve relevant literature for a specific case. The evaluated methods fall into three categories: word-vector based, Transformer based, and citation based. They have created the benchmark datasets specifically for this study. Literature recommendations are considered correct if they cover the same topic or provide essential background information for the case at hand. Specifically, a recommendation is deemed accurate if the suggested case is found in the same case-book or within the same category as the case being examined. The authors used Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) as evaluation metrics. Notably, Poincaré embeddings were generated in hyperbolic space rather than Euclidean space. The study confirms that due to the token length limitations of Transformers, they may not always be the most effective choice. The best results in both datasets were achieved through a hybrid combination of FastTextLegal and Poincaré.

In P34 [78], the authors introduced a dataset for NER specifically tailored to decisions from the German Federal Court. The dataset comprises approximately 67,000 sentences with over 2 million tokens. It includes 54,000 manually annotated entities, categorized into 19 detailed semantic classes. These classes are: person, judge, lawyer, landscape, street, city, country, organization, company, institution, court, brand, law, ordinance, European legal norm, contract, regulation, court decision, and legal literature. Remarkably, this comprehensive dataset draws upon the content of seven distinct datasets derived from seven different courts. Evaluation of the dataset employs two prominent models: Conditional Random Fields (CRF) and Bidirectional Long Short-Term Memory (Bi-LSTM).

In P35 [79], a novel dataset, known as the Contract Understanding Atticus dataset (CUAD), is introduced. CUAD is meticulously curated for legal contract review and was collaboratively developed by dozens of legal domain experts associated with the Atticus Project. The core task involves annotating salient sections of legal contracts, with the ultimate objective of reducing the time and costs typically associated with performing this task by one or more domain experts. This dataset encompasses more than 500 contracts and boasts over 13,000 annotations spanning across 41 labels or categories. Notably, CUAD stands out as one of the relatively few machine learning datasets that have been diligently curated by domain experts. The experiments were carried out exclusively using Transformer-based architectures, among which include BERT, RoBERTa, AlBERT, and DeBERTa. DeBERTa and RoBERTa emerged as the top-performing models, yet with significant room for future enhancements, as for the case of performance imbalance for each category.

In P37 [75], the authors present MAUD, a dataset based on the 2021 American bar examination. The paper introduces the concept of Merger Agreements, which are legal documents governing corporate acquisitions. The dataset comprises 47,457 annotations derived from legal texts extracted from 152 English-language merger agreements. MAUD serves as a multiple-choice reading comprehension task, with performance evaluation conducted using the Area Under the Precision and Recall Curve (AUPR) metric. Notably, this study demonstrates that larger and more recent architectural models yield improved results on the proposed task.

We can conclude that datasets and benchmarking are crucial for advancing AI technologies in the legal domain. They enable the improvement of the performance of AI models on various legal tasks. Studies have introduced important datasets like LexGLUE for NLP tasks, CUAD for contract review, and MAUD for merger agreements. These efforts underscore the importance of curated datasets and benchmarking in enhancing the capabilities of legal AI across different types of legal documents and tasks. However, it is important to note the limitations and diversity issues within these datasets. While datasets like CUAD, MAUD, and E-NER have been introduced, their generalizability across different legal subdomains remains a challenge. These datasets often reflect the specific legal contexts and jurisdictions in which they were created, which may limit their applicability to other legal systems. For instance, datasets focused on common law jurisdictions may not generalize well to civil law systems due to differences in legal frameworks and terminology. Additionally, the diversity within these datasets can be limited, as they may not fully represent the wide range of legal scenarios and demographics encountered in real-world applications. Addressing these limitations requires ongoing efforts to create more diverse and representative datasets that can better generalize across different legal subdomains.

6) DOCUMENT AUTOMATION (T9)

The integration of AI in document review and automation represents a significant advancement in the legal industry, improving efficiency, accuracy, and scalability. By leveraging AI technologies, legal professionals can handle complex document review tasks more effectively, allowing them to focus on higher-value work and deliver better outcomes for their clients [51]. We have found 5 studies that focused on this task (P17-P19, P25, P35).

The contract review process is a highly time-consuming and costly task. Small companies often sign contracts without an additional review stage. It is within this context that P17 [99] addresses the topic of contract review, with a specific focus on evaluating the fairness of a contract. They have also released a corpus of 607 annotated contracts. In their work, the authors demonstrate that certain linguistic features, such as negations, significantly complicate the task. Also, they intend to determine, framing it as an NLI task, whether a given hypothesis implies, contradicts, or is not mentioned in relation to the entire text of the contract. In this context, the task can be broken down as

follows: a) NLI that involves sentence classification for the hypothesis concerning the three mentioned classes: entailed, contradiction, not mentioned; and b) Evidence Identification that involves identifying the text span associated with the hypothesis as mentioned above (in the case of entailed or contradiction). The proposed final model (SPAN NLI BERT) achieves significantly better performance than all the considered baselines.

In P18 [100], the authors focus on detecting two crucial elements in contract reviews: entities and red flags, which are terms or sentences that indicate that there is some danger for one or more of the signing parties. Their focus is solely on leasing contracts. The released benchmark comprises 179 documents. Additionally, a new Language Model tailored for this task, called ALeaseBERT, is introduced.

In P19 [101] and in P20 [65], AI, including generative models, is employed to assist users in contract drafting. In P19, the public dataset used for this purpose is LEDGAR [101]. LEDGAR is a multilabel corpus containing legal provisions from contracts. It was created by collecting and extracting data from publicly available SEC filings, and, to the best of the authors' knowledge, is the first open-access corpus of this nature. As the dataset was developed using a semi-automated approach, several noise reduction methods were employed and are thoroughly analysed. With more than 12,000 distinct labels applied to nearly 100,000 provisions across over 60,000 contracts, LEDGAR offers significant potential for advancing research in Legal Natural Language Processing (NLP), particularly in large-scale or extreme text classification, and for supporting legal studies. The proposed approach consists of three steps. Working at the keyphrase level, rather than individual words, the authors aim to capture the semantics of broader concepts, not just isolated words. They also discuss Vicuna (based on LLAMA), which achieves 90% of ChatGPT's performance with only 7% of the total parameters. To control hallucination phenomena, the authors explore two possibilities involving human intervention: a) direct clause revision, and b) reformulating the input provided to steer a new text generation. The framework proceeds through the following three steps: 1) Sentence Transformer, representing sentences with vectors; 2) Pattern Rank, finding keyphrases most similar to the input; and 3) UMAP used for vector dimension reduction. The dataset comprises 60,540 contracts with a total of 846,274 clauses. Chat-GPT emerges as the top-performing model, although Vicuna, with far fewer parameters, still delivers notable results. In P20 [65] is proposed a model to assist in drafting legal documents by identifying regularities within paragraphs, using document embeddings and nearest neighbour search, evaluated on the EUR-Lex dataset.

In conclusion, these studies demonstrate the potential of AI-based solutions to improve the efficiency and accuracy of contract review and drafting processes. P17 [99] demonstrates the potential of NLI-based approaches to enhance contract fairness evaluation, while P18 [100] showcases a specialized language model for identifying critical elements in lease agreements. P19 [101] and P20 [65] illustrate the effectiveness of generative models and document embedding techniques in contract drafting and the identification of regularities within legal documents. These findings underscore the potential for AI-driven solutions to streamline the contract review process, making it more efficient and accessible, particularly for small companies. However, certain linguistic features, such as negations, can complicate the task, and further research is needed to address these challenges. Additionally, while AI models such as Chat-GPT and Vicuna show promising results, human intervention is still necessary to control hallucination phenomena and ensure the fairness and legal-compliances of contracts.

C. LEGAL PREDICTION

1) JUDGEMENT PREDICTION (T10)

Judgement prediction refers to forecasting the outcomes of legal cases which involves analysing vast amounts of legal data, including past court decisions, case facts, and applicable laws, to predict the verdicts of ongoing or future cases. The primary purpose of judgement prediction is to provide insights that can assist legal professionals in case strategy, risk assessment, and decision-making [19], [20], [21], [22], [52], [53]. We identified six studies focused on judgment prediction (P02, P03, P32, P50, P52), and only one for the Court Judgement Prediction and Explanation task (P01).

P01 [22] introduced the Indian Legal Documents Corpus (ILDC), a repository of Indian legal documents featuring 35,000 annotated Supreme Court cases. The ILDC serves as the foundation for the Court Judgement Prediction and Explanation (CJPE) task. The CJPE task involves predicting the final judgement based on case facts and arguments. The cases span from 1947 to April 2020 and have been examined using classical machine learning models, such as word and sentence-level embeddings, logistic regressors, SVMs, and Random Forest. Additionally, sequential models like Transformers and hierarchical Transformers have been employed, with XLNet and BiGru emerging as the two most effective models. The most proficient model achieved an accuracy rate of 78% when compared to a domain expert, who attained an accuracy of 94% on the same corpus.

Similarly, P02 [90] introduced a multilingual dataset centred around cases from the Swiss Federal Supreme Court, comprising over 85,000 cases. The authors employed state-of-the-art BERT-based models, including those surpassing the 512-token limitation. The original available case judgements, considered as labels, encompass: Approval, Partial Approval, Dismissal, Partial Dismissal, Inadmissible and Write Off. The initial four labels pertain to merit, while the last two are based on formal reasons. The models utilized encompass Standard BERT, LongBERT, and Hierarchical BERT. Results have underscored that, given the dataset's inherent imbalance, the Majority system tends to perform favourably concerning micro-F1. However, when considering Macro-F1, Hierarchical BERT emerged as the superior choice.

Another language-specific model is presented in P32 [52], which introduced a Romanian BERT model pre-trained on a large specialized corpus. The authors claimed that their model outperforms several strong baselines for legal judgement prediction on two different corpora, consisting of cases from trials involving banks in Romania. Also in this study, the beneficial effect of the domain-specific pre-training is supported by the results.

In another domain, P03 [89] presented LBOX OPEN, which is a large-scale benchmark of Korean legal AI datasets. The legal judgment prediction tasks include 10,500 criminal cases, where the model predicts fine amounts and imprisonment types based on the facts, and 4,700 civil cases, where the model predicts the degree of claim acceptance from the facts and claims for relief. The authors introduced LCUBE, an LLM pretrained on the newly created corpus, based on the GPT-2 architecture. The authors highlight their result, showing that for more difficult tasks, pre-training from scratch is more helpful than domain adaptation using fine-tuning.

Moreover, P50 [102] concentrated on predicting cases with negative outcomes instead of predicting only the case of positive outcomes. The authors made their entire codebase available on GitHub and introduced two probabilistic models to address this challenge. The set of random variables used relates to the article associated with a positive, negative, or null outcome. Another random variable encodes whether a particular article has been cited or not, and a random variable is employed to represent the textual description of the facts. The first model they propose (Joint Model) assumes that the two former random variables are conditionally independent regarding the i-th article. Furthermore, both variables solely depend on the facts. The second model presented (Claim-Outcome Model) is based on the probability that, given the facts, articles are chosen first, followed by the prediction of a specific outcome. The ECHR was used as the validation corpus, utilizing the golden labels provided in [50]. The outcome corpus was generated from this source. The results show that while a basic BERT-based classification model can predict positive outcomes with an F1 score of 75.06, it only achieves an F1 score of 10.09 for negative outcomes, performing below a random baseline, which reaches an F1 score of 11.12.

Finally, P52 [103] implemented four ML models for sentence prediction on the dataset of the ECHR. The authors investigated the impacts on performance concerning a) metrics, b) the inclusion of various combinations of parts of the considered case, c) the effect of more or less domain-specialized architectures, and d) the temporal effect of past decisions available. The dataset has been occasionally adapted for the Article Classification and Binary Classification tasks. For the former, there are 9 DSs for each article. The evaluation metrics used are Accuracy and MCC. The authors noted that the use of MCC as a metric significantly reframes the expectations and actual results of state-of-the-art models applied to the legal domain. The results showed that the Facts section plays a key role in Court Case Predictions. Also, a generalist model trained on all articles performs better than a specialized ensemble model. Additionally, predicting future events based on past cases is more difficult than using a mix of past and future cases for training.

In conclusion, with regard to this task, there are still challenges to be addressed, such as the inherent imbalance in legal datasets and the difficulty in predicting negative outcomes. Especially in this area, the language-specific trained models have proved to be the elective choice for state-of-the-art results in judgement prediction.

2) NEXT SENTENCE PREDICTION (T11)

Next Sentence Prediction using Transformer models like BERT in legal AI offers promising advancements in automating and enhancing the understanding of legal texts. By leveraging the capabilities of these models, legal professionals can improve the efficiency and accuracy of legal document drafting, research, and analysis. With regard to next sentence prediction, we have identified 4 studies that contribute to this task (P39, P41, P42, P48), addressing Miscellaneous, Legal Research, Legal Documents and Case Law, respectively (see Table 2).

In P39 [82], the authors introduce BureauBERTo, an enhanced version of the Italian Transformer model UmBERTo. By performing additional pre-training using a corpus of documents related to public administration, banking, and insurance, they expanded UmBERTo's base vocabulary. This training methodology underscores the unique characteristics of Italian administrative jargon, which extensively uses domain-specific terms (e.g., *"ravvedimento operoso"* and *"imponibile"*). Leveraging databases from the SEMPLICE[9] and ABI2LE[10] projects, the additional pre-training improved performance on proposed tasks, demonstrating BureauBERTo's efficacy for legal applications in the public administration sector.

Also in the Italian legal domain, P41 [104] addresses the limitations of using Transformers pre-trained on general corpora like Wikipedia. Given the cryptic nature of legal language and the prevalence of Latin-based terms and archaic terminology, the authors performed additional pre-training on the Italian Civil Code. Starting with ITALIAN XXL BERT, which had been pre-trained on a large corpus of 81 GB, they further trained the model using the National Jurisprudential Archive, containing millions of legal documents. This pre-training resulted in the proposed ITALIAN-LEGAL-BERT outperforming ITALIAN-BERT by 18.2% in civil cases and 15.4% in criminal cases, highlighting the importance of domain-specific pre-training for legal tasks.

[9]SEMantic instruments for PubLIc administrators and CitizEns: www.semplicepa.it (Accessed: Dec. 28, 2024)

[10]https://www.01s.it/capofila-del-progetto-abi2le-ability-to-learning/ (Accessed: Dec. 28, 2024)

P48 [105] presents ConfliBERT, where the authors demonstrate the superiority of both pre-training from scratch and continual pre-training over standard BERT. ConfliBERT was implemented using two methods, Cont and SCR, each with cased and uncased versions. The model architecture mirrors BERT-Base, with 12 layers and 110 million parameters. By employing domain-specific vocabulary (ConfliVocab) for SCR models and optimizing the learning process without the Next Sentence Prediction (NSP) task, the authors achieved enhanced performance. Pre-training on a corpus of 7 billion words using four V-100 GPUs, they found that both approaches significantly improved masked language model loss optimization, underscoring the effectiveness of tailored pre-training strategies for specialized domains.

In conclusion, these studies highlight that the extension of the vocabulary through additional pre-training on specific administrative domains, leads to significant performance improvements over general-purpose models, emphasizing the necessity of domain-specific pre-training for legal language processing. Finally, both pre-training from scratch and continual pre-training methodologies successfully enhanced performance in the legal domain, demonstrating the critical role of using a domain-specific vocabulary and optimized training strategies in achieving state-of-the-art results.

VI. TECHNOLOGIES: TRENDS WITHIN LEGAL TASKS AND REAL-LIFE IMPLEMENTATIONS

A. TRENDS

As part of the analysis and across various legal NLP tasks, we identified that Transformer models such as BERT, LEGAL-BERT, RoBERTa, and their variants are the predominant technologies driving advancements. These models are extensively used for tasks including judgement prediction, document retrieval, classification, and NER. A table to summarize the content of this section is shown in the Table 5.

There is a noticeable trend towards domain-specific adaptations of these Transformers, like LEGAL-BERT and LamBERTa, which are fine-tuned to handle the unique complexities of legal language and documents. Additionally, the integration of traditional models and embeddings, such as SVM, BM25, GloVe, and Doc2Vec, with Transformers indicates a hybrid approach where conventional methods complement the capabilities of newer, more sophisticated models. Another significant trend is the increasing use of LLMs like ChatGPT and Vicuna for tasks such as contract review and automation. This demonstrates the evolving landscape where these powerful models are being leveraged to handle more complex and nuanced legal tasks, providing more comprehensive and contextually accurate outputs.

The development and benchmarking of multilingual and multi-label datasets also highlight the growing emphasis on creating more inclusive and versatile models that can perform across different legal systems and languages. Overall, the integration of advanced Transformer models with traditional techniques and the adaptation of LLMs for specific legal tasks are shaping the future of legal NLP technologies.

In particular, in *Document Retrieval*, authors have been combining traditional retrieval methods with Transformers and integrating similarity detection techniques. Key technologies employed include BERT, Sentence-BERT, LEGAL-BERT, BM25, GloVe, Doc2Vec, CatBoost, TF-IDF, and Longformer (e.g., P04, P06, P10). This hybrid approach leverages the strengths of both traditional and modern methods to enhance the accuracy and relevance of retrieved documents [106].

For *Question Answering*, BM25, Legal GloVe, and Legal Siamese BERT are commonly used to develop systems tailored for legal queries, leveraging embeddings and traditional retrieval methods to improve performance (e.g., P49, P53). These techniques aim to accurately understand and respond to complex legal questions.

Regarding *Named Entity Recognition*, combining sequence models with Transformers and performing domain adaptation for specific legal systems has been proved to be effective. Technologies like BERT, LEGAL-BERT, RoBERTa, DeBERTa, Bi-LSTM, CRF, BureauBERTo, and UmBERTo are employed to enhance entity recognition in legal texts (e.g., P34, P35, P36).

Moreover, *Similarity Detection* has seen the use of enhanced embeddings for legal text and the integration of text and network-based similarity measures. Frequently used technologies include BERT, LEGAL-BERT, RoBERTa, Sentence-BERT, GloVe, Doc2Vec, and TF-IDF (e.g., P09, P11, P33). These methods improve the identification of similar legal documents and cases.

Also, in *Classification* extensive use of Transformer models, along with traditional classifiers for baseline comparisons, is evident. Popular technologies are BERT, LEGAL-BERT, RoBERTa, DeBERTa, Longformer, SVM, CNN, LSTM, and Bi-LSTM (e.g., P21, P26, P55). This combination ensures robust classification performance across various legal texts.

For *Document Summarization* BERT, PEGASUS, BART, LED, and PRIMERA are utilized for both extractive and abstractive summarization techniques, along with multitask learning approaches (e.g., P03, P60, P59). These technologies help generate concise and informative summaries of lengthy legal documents.

Datasets and Benchmarking efforts have focused on creating multilingual and multi-label datasets, evaluating pre-trained models on legal data. Technologies such as BERT, RoBERTa, DeBERTa, Longformer, MiniLM, DistilBERT, mDeBERTa-V3, and XML-R are frequently used (e.g., P21, P22, P23). These datasets and benchmarks are crucial for advancing legal NLP research.

Likewise, in *Document Review and Automation* leveraging LLMs models for contract drafting and automation, along with creating domain-specific BERT extensions, is common. Technologies like Vicuna, ChatGPT, Span NLI BERT, and

TABLE 5. Technologies and trends in legal NLP tasks.

Legal Task	Key Technologies	Trends and Approaches
Document Retrieval	BERT, Sentence-BERT, LEGAL-BERT, BM25, GloVe, Doc2Vec, CatBoost, TF-IDF, Longformer	Combines traditional retrieval methods with Transformers; uses similarity detection techniques.
Question Answering	BM25, Legal GloVe, Legal Siamese BERT	Tailored for legal queries; integrates embeddings and traditional retrieval methods to enhance performance.
Named Entity Recognition (NER)	BERT, LEGAL-BERT, RoBERTa, DeBERTa, Bi-LSTM, CRF, BureauBERTo, UmBERTo	Combines sequence models with Transformers; adapts models for specific legal domains.
Similarity Detection	BERT, LEGAL-BERT, RoBERTa, Sentence-BERT, GloVe, Doc2Vec, TF-IDF	Uses enhanced embeddings and integrates text and network-based similarity measures.
Classification	BERT, LEGAL-BERT, RoBERTa, DeBERTa, Longformer, SVM, CNN, LSTM, Bi-LSTM	Extensive use of Transformers; traditional classifiers for baseline comparisons.
Document Summarization	BERT, PEGASUS, BART, LED, PRIMERA	Applies both extractive and abstractive summarization; multitask learning approaches.
Datasets and Benchmarking	BERT, RoBERTa, DeBERTa, Longformer, MiniLM, DistilBERT, mDeBERTa-V3, XML-R	Focus on multilingual and multi-label datasets; evaluates pre-trained models on legal data.
Document Review and Automation	Vicuna, ChatGPT, Span NLI BERT, ALeaseBERT	LLMs used for contract drafting, review, and automation.
Judgement Prediction	BERT, LEGAL-BERT, BiGRU, RoBERTa, Logistic Regression, SVM, Random Forest, XL-Net, Longformer	Hierarchical models, domain-specific BERT models, and multilingual capabilities for predicting legal judgments.

ALeaseBERT are utilized to streamline and enhance the automation of legal document review (e.g., P17, P19).

In *Judgement Prediction* the most used technologies are BERT, LEGAL-BERT, BiGRU, RoBERTa, Logistic Regression, SVM, Random Forest, XL-Net, and Longformer. Authors employ hierarchical models, domain-specific BERT models, and multilingual capabilities to predict legal judgments effectively (e.g., P01, P02, P03). These technologies enable more accurate and nuanced predictions in judicial contexts.

We can conclude that Transformer models such as BERT, LegalBERT, RoBERTa, and their variants dominate the field, while traditional models like SVM, BM25, GloVe, and Doc2Vec are still used alongside newer technologies. Domain-specific Transformer adaptations (e.g., LegalBERT, LamBERTa) highlight the importance of fine-tuning for legal tasks. The integration of large language models (e.g., ChatGPT, Vicuna) for contract review and automation is an emerging trend, reflecting the evolving landscape of legal NLP research.

B. REAL-LIFE IMPLEMENTATIONS

To provide a practical context for our survey, we discuss several real-life implementations of LLMs in the legal domain. These examples demonstrate the feasibility and impact of LLMs in enhancing legal practice and decision-making.

1) CONTRACT REVIEW AND ANALYSIS

- **ROSS Intelligence.**[11] ROSS Intelligence is an AI-powered legal research tool that uses NLP and machine learning to assist lawyers in legal research and contract review. The platform helps legal professionals find relevant case law, analyze contracts, and identify key clauses, significantly reducing the time and effort required for these tasks.
- **Kira Systems.**[12] Kira Systems offers an AI solution for contract analysis that uses machine learning to extract and analyze key information from contracts. The tool has been implemented by numerous law firms and corporate legal departments to streamline contract review processes and improve accuracy.

2) LEGAL DOCUMENT ANALYSIS

- **LexisNexis.**[13] LexisNexis has integrated AI and NLP technologies into its legal research platform to enhance document analysis and retrieval. The platform uses LLMs to understand and categorize legal documents, making it easier for legal professionals to find relevant information and precedents.
- **Westlaw Edge.**[14] Westlaw Edge, a product of Thomson Reuters, employs AI to provide advanced legal research capabilities. The platform uses NLP to analyze legal documents, identify key issues, and predict case outcomes, helping lawyers make more informed decisions.

3) CASE PREDICTION AND OUTCOME ANALYSIS

- **CaseCrunch**[15] (now Luminance). CaseCrunch, now known as Luminance, developed an AI system for predicting the outcomes of legal cases. The platform uses machine learning algorithms to analyze case data and provide predictions, assisting lawyers in assessing the strength of their cases and developing strategies.
- **Premonition.**[16] Premonition is an AI-powered litigation analytics tool that predicts case outcomes based on historical data. The platform has been used by law firms and corporations to assess litigation risks and make data-driven decisions.

[11] https://www.rossintelligence.com/ (Accessed: Dec. 28, 2024)

[12] https://kirasystems.com/ (Accessed: Dec. 28, 2024)

[13] https://www.lexisnexis.com/ (Accessed: Dec. 28, 2024)

[14] https://legal.thomsonreuters.com/en/products/westlaw (Accessed: Dec. 28, 2024)

[15] https://www.luminance.com/ (Accessed: Dec. 28, 2024)

[16] https://www.premonition.ai/ (Accessed: Dec. 28, 2024)

4) REGULATORY COMPLIANCE

- **Ascent.**[17] Ascent uses AI to automate regulatory compliance for financial institutions. The platform employs NLP to ananalyseegulatory texts and identify relevant obligations, helping organizations stay compliant with complex and ever-changing regulations.
- **Compliance.ai.**[18] Compliance.ai offers an AI-driven regulatory compliance solution that uses NLP to monitor and analyse regulatory changes. The platform helps organizations track and comply with regulatory requirements, reducing the risk of non-compliance.

These real-life implementations represent practical applications and benefits of LLMs in the legal domain. By highlighting these examples, we aim to provide a comprehensive understanding of the current state and potential of LLMs in enhancing legal practice and decision-making.

VII. CONCLUSION

In this work, we have explored recent applications of LLMs models to the legal domain. After performing a detailed review of the recent literature, we formulated four research questions. According to all the works discussed in the previous sections, here we conclude our work listing our main findings and mentioning the specific RQs formulated in the Section I and answered here.

A. (RQ1) WHAT ARE THE NLP APPLICATIONS IN THE LEGAL DOMAIN THAT HAVE ALREADY MADE USE OF LLMS?

1) LEGAL SEARCH IS A CRITICAL COMPONENT OF LEGAL RESEARCH AND PRACTICE

The studies highlight the importance of document retrieval, case entailment and question answering in facilitating legal research, case preparation, document drafting, and client advice. The legal domain is characterized by a vast amount of unstructured and semi-structured information, making efficient document retrieval crucial for legal professionals to make informed decisions. The majority of the studies are mainly focused on case law retrieval, demonstrating its significance in the legal domain. Case law retrieval involves identifying relevant precedents and previous judgments to inform current legal decisions. The complexity of this task lies in capturing the nuances of legal language and the context in which legal cases are presented.

2) REAL-WORLD DATASETS ARE REQUIRED FOR TRAINING AND EVALUATING LEGAL DOCUMENT RETRIEVAL MODELS

The studies relied on real-world datasets, such as LexGLUE, CaseHOLD, and EUR-Lex, to train and evaluate their models. Real-world datasets provide a realistic representation of the complexities and nuances of legal language, enabling models to learn from practical applications. Datasets and benchmarking enable the assessment of model performance to improve the performance of AI tools on various legal tasks. Studies have introduced important datasets like LexGLUE for NLP tasks, CUAD for contract review, and MAUD for merger agreements. These efforts underscore the importance of curated datasets and benchmarking in enhancing the capabilities of legal AI across different types of legal documents and tasks.

B. (RQ2) IS THE PRE-TRAINING OR THE FINE-TUNING OF LLMS SPECIFICALLY FOR THE LEGAL DOMAIN ALWAYS BENEFICIAL? AND IN WHAT CASES AND FOR WHAT TASKS?

1) DOCUMENT EMBEDDING MODELS ARE OFTEN EMPLOYED FOR CAPTURING SEMANTIC RELATIONSHIPS IN LEGAL DOCUMENTS

The studies highlighted the importance of document embedding models in capturing semantic relationships between documents and retrieving relevant information. These models can learn to represent legal documents as dense vectors, enabling efficient comparison and retrieval in terms of distance in vector space. According to several studies, word embeddings trained from scratch are able to outperform pre-trained embedding and traditional approaches as TF-IDF. Thanks to the word embedding, the studies highlight the potential of text similarity, document clustering, and topic modelling techniques for legal document analysis. The proposed methods have shown promising results in improving the performance of legal document analysis tasks, such as law article retrieval, case retrieval, and topic modelling.

2) TASK-SPECIFIC MODELS USUALLY OUTPERFORM GENERAL-DOMAIN PRE-TRAINED MODELS ON LEGAL DOCUMENT RETRIEVAL TASKS

Legal texts often contain synonyms and specialized jargon, which can affect similarity results. The current approaches include several mechanisms to handle these aspects effectively. Pre-trained language models like BERT, Sentence-BERT, and LEGAL-BERT capture semantic relationships and understand context, helping to recognize synonyms and interpret jargon accurately. Domain-specific fine-tuning further enhances these models' ability to understand legal terms. Embedding techniques such as GloVe and Doc2Vec generate word and document embeddings that capture semantic similarities. Additionally, some approaches incorporate legal synonym dictionaries and ontologies to map different terms to their standardized forms. Contextual analysis, combined with models like BM25 and TF-IDF, helps in handling the variability in legal language by considering the context in which terms appear. The studies emphasized the importance of domain adaptation for legal tasks, including fine-tuning pre-trained models on legal datasets or using domain-specific knowledge graphs. This approach enables models to learn from the specific characteristics of legal language and adapt to the unique challenges of the legal domain.

[17] https://ascent.ai/ (Accessed: Dec. 28, 2024)
[18] https://www.compliance.ai/ (Accessed: Dec. 28, 2024)

Several studies demonstrated that task-specific models, such as SAILER, LEGAL-BERT and LamBERTa, can outperform general-domain pre-trained models like BERT on specific tasks like legal case retrieval. This suggests that domain adaptation is crucial, as task-specific models can learn to capture the nuances of legal language and context. Even if some tasks (e.g. entailment) are not yet fully solved in the literature. LLM-based approaches have consistently outperformed traditional and statistical method, due to their ability to comprehend contextual and semantic relations between words.

C. (RQ3) WHAT ARE THE MAIN ADVANTAGES AND LIMITATIONS OF LLMS IN THE LEGAL FIELD?

Long-Document Retrieval Presents a Unique Challenge in Legal Document Retrieval. Despite progress made in legal document retrieval, there are still several open challenges that require further research attention. These include handling ambiguity in legal language, dealing with irrelevant or redundant information, and improving model interpretability. The increasing complexity of legal documents and the need for efficient retrieval methods highlight the importance of developing novel approaches to long-document retrieval. This may involve leveraging advancements in NLP, information retrieval, or machine learning techniques to overcome the limitations. Five studies specifically addressed long-document retrieval, emphasizing the need for efficient and effective methods to retrieve relevant information from lengthy documents. Long documents often require novel approaches to capture the semantic relationships between documents and retrieve relevant information. Unfortunately, it is evident from the literature [107] that performance might suffer greatly when essential information is moved within long context, suggesting that language models as they exist today are not able to effectively utilize information in lengthy input scenarios. The results indicate that there is still room for improvement in terms of summary quality and accuracy, particularly in low-resource environments and for long legal documents. The use of domain-specific pre-trained models, such as LEGAL-BERT, and the creation of new datasets, such as ITA-CaseHold and Multi-LexSum, are important steps towards improving the performance of NLP techniques in summarizing legal documents. However, they do not yet fully solve the issues related to the understanding of long documents (RQ3).

D. (RQ4) WHAT ARE THE POSSIBLE APPLICATIONS OF LLMS TO THE LEGAL DOMAIN NOT YET FULLY EXPLORED?

Finally, there is a growing interest in the field of prompt engineering for interacting with LLMs [7], [108], [109] (RQ4). We argue that future applications will be mainly based on human-centered techniques [110], [111]. All of these techniques are usually based on zero-shot and few-shot prompting, and more generically on the concept of In-Context Learning (ICL) [112]. In the case of LLM, ICL refers to the model's ability to generate answers or perform tasks based on examples provided within the input prompt, without updating its internal parameters, by leveraging the contextual information to generate appropriate and desired outputs. Crucially, there's no need to adjust the model parameters for the new assignment. The work on pre-trained big language models, which are capable of performing ICL without requiring retraining, has popularized ICL. It may be expected that similar approaches will also be further employed in the legal domain to interact in a Q&A manner with the LLMs to accomplish legal tasks without further pre-training.

VIII. FUTURE DIRECTIONS

Based on the findings and analysis presented in our survey paper on the application of LLMs in the legal domain, in this section we propose the following future directions and suggestions to advance the field.

A. ENHANCED DATASET DIVERSITY AND QUALITY

Development of Multilingual Datasets. Future research should focus on creating and curating multilingual datasets to support the development of LLMs that can handle legal texts in various languages. This will enhance the applicability of these models in global legal contexts. **Inclusion of Diverse Legal Subdomains**. Efforts should be made to include datasets from diverse legal subdomains, such as intellectual property, environmental law, and international law, to ensure that LLMs can generalize well across different areas of law.

B. ADVANCED MODEL ARCHITECTURES

Hybrid Models. Explore the integration of LLMs with other AI techniques, such as rule-based systems and knowledge graphs, to create hybrid models that leverage the strengths of different approaches. **Domain-Specific Fine-Tuning**. Further investigate advanced fine-tuning techniques tailored to specific legal tasks, such as contract analysis, legal research, and case outcome prediction, to improve the performance and accuracy of LLMs in these areas.

C. ETHICAL AND LEGAL CONSIDERATIONS

Bias and Fairness. Conduct studies to identify and mitigate biases in LLMs used for legal tasks. This includes developing methods to ensure fairness and transparency in model predictions and decisions. **Data Privacy and Security**. Address data privacy and security concerns associated with the use of LLMs in legal applications. This may involve developing secure data handling protocols and ensuring compliance with relevant data protection regulations.

D. INTERDISCIPLINARY COLLABORATION

Legal Expertise Integration. Foster collaboration between AI researchers and legal experts to ensure that the development and deployment of LLMs in the legal domain are informed by practical legal knowledge and insights. **Cross-Disciplinary Research**. Encourage research that bridges the

gap between AI and other disciplines, such as ethics, sociology, and psychology, to address the broader implications of using LLMs in legal contexts.

E. STANDARDIZED BENCHMARKS AND METRICS

Develop and adopt standardized benchmarks and evaluation metrics specifically tailored for legal tasks. This will facilitate fair and consistent comparisons of different LLMs and their performance in legal applications.

F. CONTINUOUS LEARNING AND UPDATING

Implement continuous learning and updating mechanisms for LLMs to keep them up-to-date with the latest legal developments and changes in legislation. This will ensure that the models remain relevant and accurate over time.

G. EDUCATION AND TRAINING

Provide education and training programs for legal professionals to enhance their understanding and effective use of LLMs in their practice. This includes workshops, courses, and resources that explain the capabilities and limitations of these models.

H. REGULATORY FRAMEWORKS

Work towards the development of regulatory frameworks that govern the use of LLMs in legal applications. This will help ensure that these models are used responsibly and ethically, while also promoting innovation and progress in the field.

AUTHOR CONTRIBUTIONS

Marco Siino and Mariana Falco were involved in investigation, conceptualization, formal analysis, methodology, validation, visualization, writing original draft, review, and editing. Daniele Croce and Paolo Rosso were involved in methodology, supervision, writing review, and editing. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] E. Mumcuolu, C. E. Öztürk, H. M. Ozaktas, and A. Ko, "Natural language processing in law: Prediction of outcomes in the higher courts of Turkey," *Inf. Process. Manage.*, vol. 580, no. 5, 2021, Art. no. 0102684.

[2] D. F. Engstrom, D. E. Ho, C. M. Sharkey, and M.-F. Cuéllar, "Government by algorithm: Artificial intelligence in federal administrative agencies," *NYU School Law, Public Law Res. Paper*, vol. 12, pp. 20–54, Mar. 2020.

[3] C. M. Greco and A. Tagarelli, "Bringing order into the realm of transformer-based language models for artificial intelligence and law," *Artif. Intell. Law*, vol. 310, no. 2, p. 148, 2023.

[4] F. B. Wiener, "Decision prediction by computers: Nonsense cubed—And worse," *Amer. Bar Assoc. J.*, vol. 10, pp. 1023–1028, Apr. 1962.

[5] F. R. Dickerson, "The electronic searching of law," *Amer. Bar Assoc. J.*, vol. 470, no. 9, pp. 902–908, 1961.

[6] R. C. Lawlor, "What computers can do: Analysis and prediction of judicial decisions," *Amer. Bar Assoc. J.*, vol. 490, no. 4, pp. 337–344, 1963.

[7] M. Siino, "T5-medical at SemEval-2024 task 2: Using T5 medical embedding for natural language inference on clinical trial data," in *Proc. 18th Int. Workshop Semantic Eval. (SemEval)*, Mexico City, Mexico, A. K. Ojha, A. S. Doğruöz, H. T. Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds., 2024, pp. 40–46.

[8] K. K. Bressem, J.-M. Papaioannou, P. Grundmann, F. Borchert, L. C. Adams, L. Liu, F. Busch, L. Xu, J. P. Loyen, S. M. Niehues, M. Augustin, L. Grosser, M. R. Makowski, H. J. W. L. Aerts, and A. Löser, "MedBERT.de: A comprehensive German BERT model for the medical domain," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121598.

[9] S. Wada, T. Takeda, K. Okada, S. Manabe, S. Konishi, J. Kamohara, and Y. Matsumura, "Oversampling effect in pretraining for bidirectional encoder representations from transformers (BERT) to localize medical BERT and enhance biomedical BERT," *Artif. Intell. Med.*, vol. 153, Jul. 2024, Art. no. 102889.

[10] M. Siino and I. Tinnirello, "GPT prompt engineering for scheduling appliances usage for energy cost optimization," in *Proc. IEEE Int. Symp. Meas. Netw.*, Rome, Italy, Jul. 2024, pp. 1–6.

[11] M. Siino, F. Giuliano, and I. Tinnirello, "LLM application for knowledge extraction from networking log files," in *Proc. 4th Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, Nov. 2024, pp. 1–6.

[12] T. Bench-Capon et al., "A history of AI and law in 50 papers: 25 years of the international conference on AI and law," *Artif. Intell. Law*, vol. 20, no. 3, pp. 215–319, Sep. 2012.

[13] I. Chalkidis and D. Kampas, "Deep learning in law: Early adaptation and legal word embeddings trained on large corpora," *Artif. Intell. Law*, vol. 27, no. 2, pp. 171–198, Jun. 2019.

[14] M. Siino, "Mistral at SemEval-2024 task 5: Mistral 7B for argument reasoning in civil procedure," in *Proc. 18th Int. Workshop Semantic Eval. (SemEval)*, Mexico City, Mexico, A. K. Ojha, A. S. Doğruöz, H. T. Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds., 2024, pp. 155–162.

[15] C. M. Greco, A. Tagarelli, and E. Zumpano, "A comparison of transformer-based language models on NLP benchmarks," in *Proc. 27th Int. Conf. Appl. Natural Lang. Inf. Syst.*, vol. 13286, Valencia, Spain, P. Rosso, V. Basile, R. Martínez, E. Métais, and F. Meziane, Eds., Cham, Switzerland: Springer, Jan. 2022, pp. 490–501.

[16] D. Aumiller, S. Almasian, S. Lackner, and M. Gertz, "Structural text segmentation of legal documents," in *Proc. 18th Int. Conf. Artif. Intell. Law*, São Paulo, Brazil, J. Maranhão and A. Z. Wyner, Eds., Jun. 2021, pp. 2–11.

[17] R. Z. Mahari, "AutoLAW: Augmented legal reasoning through legal precedent prediction," 2021, *arXiv:2106.16034*.

[18] J. Rabelo, R. Goebel, M.-Y. Kim, Y. Kano, M. Yoshioka, and K. Satoh, "Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021," *Rev. Socionetw. Strategies*, vol. 16, no. 1, pp. 111–133, Apr. 2022.

[19] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European court of human rights: A natural language processing perspective," *PeerJ Comput. Sci.*, vol. 2, p. e93, Oct. 2016.

[20] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, "Legal judgment prediction via topological learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsuji, Eds., 2018, pp. 3540–3549.

[21] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., 2019, pp. 4317–4323.

[22] V. Malik, R. Sanjay, S. K. Nigam, K. Ghosh, S. K. Guha, A. Bhattacharya, and A. Modi, "ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., 2021, pp. 4046–4062.

[23] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help?: Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings," in *Proc. 18th Int. Conf. Artif. Intell. Law*, São Paulo, Brazil, J. Maranhão and A. Z. Wyner, Eds., Jun. 2021, pp. 159–168.

[24] N. Limsopatham, "Effectively leveraging BERT for legal document classification," in *Proc. Natural Legal Lang. Process. Workshop*, Punta Cana, Dominican Republic, N. Aletras, I. Androutsopoulos, L. Barrett, C. Goanta, and D. Preotiuc-Pietro, Eds., 2021, pp. 210–216.

[25] J. Lam, D. Liang, S. Dahan, and F. Zulkernine, "The gap between deep learning and law: Predicting employment notice," in *Proc. Natural Legal Lang. Process. Workshop*, vol. 2645, N. Aletras, I. Androutsopoulos, L. Barrett, A. Meyers, and D. Preotiuc-Pietro, Eds., 2020, pp. 52–56.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., Jun. 2017, pp. 5998–6008.
[27] J. S. Melton and R. C. Bensing, "Searching legal literature electronically: Results of a test program," *Minnesota Law Rev.*, vol. 450, no. 2, pp. 229–248, Jan. 1960.
[28] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit legal system: A summary of legal artificial intelligence," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds., 2020, pp. 5218–5230.
[29] C. Sansone and G. Sperlí, "Legal information retrieval systems: State-of-the-art and open issues," *Inf. Syst.*, vol. 106, May 2022, Art. no. 101967.
[30] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, and M. J. Bommarito, "Natural language processing in the legal domain," 2023, *arXiv:2302.12039*.
[31] R. Dale, "Law and word order: NLP in legal tech," *Natural Lang. Eng.*, vol. 25, no. 1, pp. 211–217, Jan. 2019.
[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., Jan. 2015, pp. 1–15.
[33] Z. Sun, "A short survey of viewing large language models in legal aspect," 2023, *arXiv:2303.09136*.
[34] D. Trautmann, A. Petrova, and F. Schilder, "Legal prompt engineering for multilingual legal judgement prediction," 2022, *arXiv:2212.02199*.
[35] A. Blair-Stanek, N. Holzenberger, and B. V. Durme, "Can GPT-3 perform statutory reasoning?" *Vanderbilt Law Rev.*, vol. 76, pp. 59–90, Jun. 2023.
[36] F. Yu, L. Quartey, and F. Schilder, "Legal prompting: Teaching a language model to think like a lawyer," 2022, *arXiv:2212.01326*.
[37] J. J. Nay, "Large language models as fiduciaries: A case study toward robustly communicating with artificial intelligence through legal standards," 2023, *arXiv:2301.10095*.
[38] K. Y. Iu and V. M.-Y. Wong, "Chatgpt by OpenAI: The end of litigation lawyers?" *SSRN*, 2023.
[39] S. Hargreaves, "'Words are flowing out like endless rain into a paper cup': ChatGPT & Law School Assessments," *Chin. Univ. Hong Kong Fac. Law Res. Paper*, vol. 33, p. 69, Apr. 2023.
[40] J. H. Choi, K. E. Hickman, A. Monahan, and D. B. Schwarcz, "ChatGPT goes to law school," *J. Legal Educ.*, vol. 71, p. 387, May 2023.
[41] T. P. Oltz, "ChatGPT, professor of law," *SSRN4347630*, 2023.
[42] R. Macey-Dare, "ChatGPT & generative AI systems as quasi-expert legal advice lawyers–case study considering potential appeal against conviction of tom Hayes," *SSRN 4342686*, 2023.
[43] S. Shaghaghian, L. Y. Feng, B. Jafarpour, and N. Pogrebnyakov, "Customizing contextualized language models for legal document reviews," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Atlanta, GA, USA, X. Wu, C. Jermaine, L. Xiong, X. Hu, O. Kotevska, S. Lu, W. Xu, S. Aluru, C. Zhai, E. Al-Masri, Z. Chen, and J. Saltz, Eds., Dec. 2020, pp. 2139–2148.
[44] M.-Y. Kim, J. Rabelo, H. K. B. Babiker, M. A. Rahman, and R. Goebel, "Legal information retrieval and entailment using transformer-based approaches," *Rev. Socionetwork Strategies*, vol. 18, no. 1, pp. 101–121, Apr. 2024.
[45] J. Martínez-Gil, "A survey on legal question–answering systems," *Comput. Sci. Rev.*, vol. 48, Mar. 2023, Art. no. 100552.
[46] M. P. Prajwal and M. Anand Kumar, "Legal text analysis using pre-trained transformers," in *Advanced Machine Intelligence and Signal Processing*, D. Gupta, K. Sambyo, M. Prasad, and S. Agarwal, Eds., Singapore: Springer, 2022, pp. 493–504.
[47] P. Bambroo and A. Awasthi, "LegalDB: Long DistilBERT for legal document classification," in *Proc. Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT)*, Feb. 2021, pp. 1–4.
[48] S. Klaus, R. Van Hecke, K. D. Naini, I. S. Altingovde, J. Bernabé-Moreno, and E. Herrera-Viedma, "Summarizing legal regulatory documents using transformers," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Madrid, Spain, E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds., Jul. 2022, pp. 2426–2430.
[49] D. Núñez-Robinson, J. Talavera-Montalto, and W. Ugarte, "A comparative analysis on the summarization of legal texts using transformer models," in *Proc. Adv. Res. Technol., Inf., Innov. Sustainability*, vol. 1675, Santiago de Compostela, Spain, T. Guarda, F. Portela, and M. F. Augusto, Eds., Cham, Switzerland: Springer, Jan. 2022, pp. 372–386.
[50] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, S. Muresan, P. Nakov, and A. Villavicencio, Eds., 2022, pp. 4310–4330.
[51] S. G. Graham, H. Soltani, and O. Isiaq, "Natural language processing for legal document review: Categorising deontic modalities in contracts," *Artif. Intell. Law*, Nov. 2023.
[52] M. Masala, R. C. A. Iacob, A. S. Uban, M. Cidota, H. Velicu, T. Rebedea, and M. Popescu, "JurBERT: A Romanian BERT model for legal judgement prediction," in *Proc. Natural Legal Lang. Process. Workshop*, Punta Cana, Dominican Republic, N. Aletras, I. Androutsopoulos, L. Barrett, C. Goanta, and D. Preotiuc-Pietro, Eds., 2021, pp. 86–94.
[53] Y. Huang, X. Shen, C. Li, J. Ge, and B. Luo, "Dependency learning for legal judgment prediction with a unified text-to-text transformer," 2021, *arXiv:2112.06370*.
[54] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. 71, Mar. 2021.
[55] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng.*, London, U.K., M. J. Shepperd, T. Hall, and I. Myrtveit, Eds., May 2014, pp. 1–10.
[56] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Minneapolis, MN, USA, J. Burstein, C. Doran, and T. Solorio, Eds., Jan. 2018, pp. 4171–4186.
[57] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," *Inf. Softw. Technol.*, vol. 50, no. 9, pp. 833–859, Aug. 2008.
[58] N. Jain and G. Goel, "An approach to get legal assistance using artificial intelligence," in *Proc. 8th Int. Conf. Rel., INFOCOM Technol. Optim. (ICRITO)*, Jun. 2020, pp. 768–771.
[59] H. Alberts, A. Ipek, R. Lucas, and P. Wozny, "COLIEE 2020: Legal information retrieval and entailment with legal embeddings and boosting," in *Proc. JSAI Int. Symp. Artif. Intell.* Cham, Switzerland: Springer, Jan. 2021, pp. 211–225.
[60] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *Proc. Int. Conf. Electr., Electron., Optim. Techn. (ICEEOT)*, Mar. 2016, pp. 61–66.
[61] F. Sovrano, M. Palmirani, and F. Vitali, "Legal knowledge extraction for knowledge graph based question-answering," in *Proc. Legal Knowl. Inf. Syst.*, in Frontiers in Artificial Intelligence and Applications, vol. 334, Brno, Czech Republic, Dec. 2020, pp. 143–153.
[62] Q. M. Bui, C. Nguyen, D.-T. Do, N.-K. Le, D.-H. Nguyen, T.-T.-T. Nguyen, M.-P. Nguyen, and M. L. Nguyen, "JNLP team: Deep learning approaches for tackling long and ambiguous legal documents in COLIEE 2022," in *New Frontiers in Artificial Intelligence*, Y. Takama, K. Yada, K. Satoh, and S. Arai, Eds., Cham, Switzerland: Springer, 2023, pp. 68–83.
[63] D. Mamakas, P. Tsotsi, I. Androutsopoulos, and I. Chalkidis, "Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer," in *Proc. Natural Legal Lang. Process. Workshop*, Abu Dhabi, United Arab Emirates, N. Aletras, I. Chalkidis, L. Barrett, C. Goanta, and D. Preotiuc-Pietro, Eds., 2022, pp. 130–142.
[64] V. Bellandi, S. Castano, P. Ceravolo, E. Damiani, A. Ferrara, S. Montanelli, S. Picascia, A. Polimeno, and D. Riva, "Knowledge-based legal document retrieval: A case study on Italian civil court decisions," in *Proc. 23rd Int. Conf. Knowl. Eng. Knowl. Manage.*, vol. 3256, Bozen-Bolzano, Italy, D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. Di Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, and N. Troquard, Eds., 2022, pp. 1–16.
[65] G. De Martino, G. Pio, and M. Ceci, "PRILJ: An efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments," *Artif. Intell. Law*, vol. 30, no. 3, pp. 359–390, Sep. 2022.
[66] S. Althammer, A. Askari, S. Verberne, and A. Hanbury, "DoSSIER@ COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval," 2021, *arXiv:2108.03937*.

[67] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, and S. Ma, "BERT-PLI: Modeling paragraph-level interactions for legal case retrieval," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, C. Bessiere, Ed., Jul. 2020, pp. 3501–3507.

[68] I. Almuslim and D. Inkpen, "Document level embeddings for identifying similar legal cases and laws," in *Proc. Work. Notes-Forum Inf. Retr. Eval. (FIRE)*, Hyderabad, India, P. Mehta, T. Mandl, P. Majumder, and M. Mitra, Eds., Jan. 2020, pp. 42–48.

[69] A. Tagarelli and A. Simeri, "Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code," *Artif. Intell. Law*, vol. 30, no. 3, pp. 417–473, Sep. 2022.

[70] A. Simeri and A. Tagarelli, "Exploring domain and task adaptation of Lamberta models for article retrieval on the Italian civil code," in *Proc. 19th Conf. Inf. Res. Sci. Connecting Digit. Library Sci.*, vol. 3365, Bari, Italy, A. Falcon, S. Ferilli, A. Bardi, S. Marchesin, and D. Redavid, Eds., 2023, pp. 130–143.

[71] H. Li, Q. Ai, J. Chen, Q. Dong, Y. Wu, Y. Liu, C. Chen, and Q. Tian, "SAILER: Structure-aware pre-trained language model for legal case retrieval," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Taipei, Taiwan, H.-H. Chen, W.-J. E. Duh, H.-H. Huang, M. P. Kato, J. Mothe, and B. Poblete, Eds., Jul. 2023, pp. 1035–1044.

[72] M. Ostendorff, E. Ash, T. Ruas, B. Gipp, J. Moreno-Schneider, and G. Rehm, "Evaluating document representations for content-based legal literature recommendations," in *Proc. 18th Int. Conf. Artif. Intell. Law*, São Paulo, Brazil, J. Maranhão and A. Z. Wyne, Eds., Jun. 2021, pp. 109–118.

[73] M. Siino, F. Lomonaco, and P. Rosso, "Backtranslate what you are saying and I will tell who you are," *Expert Syst.*, vol. 41, no. 8, Aug. 2024, Art. no. e13568.

[74] F. Lomonaco, M. Siino, and M. Tesconi, "Text enrichment with Japanese language to profile cryptocurrency influencers," in *Proc. Work. Notes Conf. Labs Eval. Forum (CLEF)*, vol. 3497, Thessaloniki, Greece, M. Aliannejadi, G. Faggioli, N. Ferro, and M. Vlachos, Eds., 2023, pp. 2708–2716.

[75] S. Wang, A. Scardigli, L. Tang, W. Chen, D. Levkin, A. Chen, S. Ball, T. Woodside, O. Zhang, and D. Hendrycks, "MAUD: An expert-annotated legal NLP dataset for merger agreement understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Singapore, H. Bouamor, J. Pino, and K. Bali, Eds., 2023, pp. 16369–16382.

[76] S. Khazaeli, J. Punuru, C. Morris, S. Sharma, B. Staub, M. Cole, S. Chiu-Webster, and D. Sakalley, "A free format legal question answering system," in *Proc. Natural Legal Lang. Process. Workshop*, Punta Cana, Dominican Republic, N. Aletras, I. Androutsopoulos, L. Barrett, C. Goanta, and D. Preotiuc-Pietro, Eds., pp. 107–113.

[77] M. Miliani, L. C. Passaro, and A. Lenci, "FRAQUE: A frame-based question-answering system for the public administration domain," in *Proc. 1st Workshop Lang. Technol. Government Public Admin.*, Marseille, France, D. Samy, D. Pérez-Fernández, and J. Arenas-García, Eds., May 2020, pp. 7–14.

[78] E. Leitner, G. Rehm, and J. M. Schneider, "A dataset of German legal documents for named entity recognition," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Jan. 2020, pp. 4478–4485.

[79] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An expert-annotated NLP dataset for legal contract review," in *Proc. Neural Inf. Process. Syst.*, J. Vanschoren and S.-K. Yeung, Eds., Jan. 2021, pp. 1–13.

[80] V. Pais, M. Mitrofan, C. L. Gasan, V. Coneschi, and A. Ianov, "Named entity recognition in the Romanian legal domain," in *Proc. Natural Legal Lang. Process. Workshop*, Punta Cana, Dominican Republic, N. Aletras, I. Androutsopoulos, L. Barrett, C. Goanta, and D. Preotiuc-Pietro, Eds., 2021, pp. 9–18.

[81] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of law school," 2020, *arXiv:2010.02559*.

[82] S. Auriemma, M. Madeddu, M. Miliani, A. Bondielli, L. C. Passaro, and A. Lenci, "BureauBERTo: Adapting UmBERTo to the Italian bureaucratic language," in *Proc. Italia Intelligenza Artificiale*, vol. 3486, Pisa, Italy, F. Falchi, F. Giannotti, A. Monreale, C. Boldrini, S. Rinzivillo, and S. Colantonio, Eds., 2023, pp. 240–248.

[83] D. Licari and G. Comandè, "ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the Italian legal domain," *Comput. Law Secur. Rev.*, vol. 52, Apr. 2024, Art. no. 105908.

[84] Y. Chen, Y. Zhang, J. Wang, and X. Zhang, "YNU-HPCC at SemEval-2023 task 6: LEGAL-BERT based hierarchical BiLSTM with CRF for rhetorical roles prediction," in *Proc. The 17th Int. Workshop Semantic Eval. (SemEval)*, Toronto, ON, Canada, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. T. Madabushi, R. Kumar, and E. Sartori, Eds., 2023, pp. 2075–2081.

[85] T. W. T. Au, V. Lampos, and I. Cox, "E-NER—An annotated named entity recognition corpus of legal text," in *Proc. Natural Legal Lang. Process. Workshop*, Abu Dhabi, United Arab Emirates, N. Aletras, I. Chalkidis, L. Barrett, C. Goanta, and D. Preotiuc-Pietro, Eds., 2022, pp. 246–255.

[86] S. Auriemma, M. Miliani, A. Bondielli, L. C. Passaro, and A. Lenci, "Evaluating pre-trained transformers on Italian administrative texts," in *Proc. of 1st Workshop AI Public Admin.*, vol. 3285, Udine, Italy, P. Lops, P. Basile, L. Siciliani, V. Taccardi, M. Di Ciano, and N. Lopane, Eds., 2022, pp. 54–70.

[87] P. Bhattacharya, K. Ghosh, A. Pal, and S. Ghosh, "Legal case document similarity: You need both network and text," *Inf. Process. Manage.*, vol. 59, no. 6, Nov. 2022, Art. no. 103069.

[88] R. Silveira, C. G. Fernandes, J. A. M. Neto, V. Furtado, and J. E. P. Filho, "Topic modelling of legal documents via LEGAL-BERT," in *Proc. 1st Int. Workshop Rel.-Relations Legal Domain*, So Paulo, Brazil, Jan. 2023, pp. 1–18.

[89] W. Hwang, D. Lee, K. Cho, H. Lee, and M. Seo, "A multi-task benchmark for Korean legal language understanding and judgement prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Jan. 2022, pp. 1–15.

[90] J. Niklaus, V. Matoshi, P. Rani, A. Galassi, M. Stürmer, and I. Chalkidis, "LEXTREME: A multi-lingual and multi-task benchmark for the legal domain," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Singapore, H. Bouamor, J. Pino, and K. Bali, Eds., 2023, pp. 3016–3054.

[91] S. Bano and S. Khalid, "BERT-based extractive text summarization of scholarly articles: A novel architecture," in *Proc. Int. Conf. Artif. Intell. Things (ICAIoT)*, Dec. 2022, pp. 1–5.

[92] S. Bano, S. Khalid, N. M. Tairan, H. Shah, and H. A. Khattak, "Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 9, Oct. 2023, Art. no. 101739.

[93] S. Vatsal, A. Meyers, and J. E. Ortega, "Classification of U.S. supreme court cases using BERT-based techniques," in *Proc. Conf. Recent Adv. Natural Lang. Process.-Large Lang. Models Natural Lang. Process.*, Varna, Bulgaria, R. Mitkov and G. Angelova, Eds., 2023, pp. 1207–1215.

[94] Z. Shen, K. Lo, L. Yu, N. Dahlberg, M. Schlanger, and D. Downey, "Multi-LexSum: Real-world summaries of civil rights lawsuits at multiple granularities," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 13158–13173.

[95] A. Agarwal, S. Xu, and M. Grabmair, "Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Abu Dhabi, United Arab Emirates, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., 2022, pp. 1857–1872.

[96] B. Clavié and M. Alphonsus, "The unreasonable effectiveness of the baseline: Discussing SVMs in legal text classification," in *Proc. Legal Knowl. Inf. Syst.-34th Annu. Conf.*, in Frontiers in Artificial Intelligence and Applications, vol. 346, Vilnius, Lithuania, Dec. 2021, pp. 58–61.

[97] P. Henderson, M. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, and D. E. Ho, "Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset," in *Proc. Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Jan. 2022, pp. 1–17.

[98] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX–A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Punta Cana, Dominican Republic, M.-F. Moens, X. Huang, L. Specia, and S. W.-T. Yih, Eds., 2021, pp. 6974–6996.

[99] Y. Koreeda and C. Manning, "ContractNLI: A dataset for document-level natural language inference for contracts," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Punta Cana, Dominican Republic, M.-F. Moens, X. Huang, L. Specia, and S. W.-T. Yih, Eds., 2021, pp. 1907–1919.

[100] S. Leivaditi, J. Rossi, and E. Kanoulas, "A benchmark for lease contract review," 2020, *arXiv:2010.10386*.

[101] D. Tuggener, P. von Däniken, T. Peetz, and M. Cieliebak, "LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., pp. 1235–1241.

[102] J. Valvoda, R. Cotterell, and S. Teufel, "On the role of negative precedent in legal outcome prediction," in *Proc. Trans. Assoc. Comput. Linguistics*, vol. 11, Jan. 2023, pp. 34–48.

[103] C. Steging, S. Renooij, and B. Verheij, "Taking the law more seriously by investigating design choices in machine learning prediction research," in *Proc. 6th Workshop Automated Semantic Anal. Inf. Legal Text*, vol. 3441, Braga, Portugal, F. Lagioia, J. Mumford, D. Odekerken, and H. Westermann, Eds., 2023, pp. 49–59.

[104] D. Licari and G. Comandè, "ITALIAN-LEGAL-BERT: A pre-trained transformer language model for Italian law," in *Proc. 23rd Int. Conf. Knowl. Eng. Knowl. Manage.*, vol. 3256, Bozen-Bolzano, Italy, D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. Di Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, and N. Troquard, Eds., 2022, pp. 1–16.

[105] Y. Hu, M. Hosseini, E. S. Parolin, J. Osorio, L. Khan, P. Brandt, and V. D'Orazio, "ConfliBERT: A pre-trained language model for political conflict and violence," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 5469–5482.

[106] S. Khalid, S. Wu, A. Alam, and I. Ullah, "Real-time feedback query expansion technique for supporting scholarly search using citation network analysis," *J. Inf. Sci.*, vol. 47, no. 1, pp. 3–15, Feb. 2021.

[107] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," in *Proc. Trans. Assoc. Comput. Linguistics*, vol. 12, Jan. 2024, pp. 157–173.

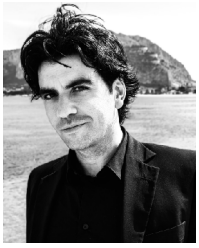
[108] M. Siino, "BadRock at SemEval-2024 task 8: DistilBERT to detect multigenerator, multidomain and multilingual black-box machine-generated text," in *Proc. 18th Int. Workshop Semantic Eval. (SemEval)*, Mexico City, Mexico, A. K. Ojha, A. S. Doğruöz, H. T. Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds., 2024, pp. 239–245.

[109] M. Siino, "McRock at SemEval-2024 task 4: Mistral 7B for multilingual detection of persuasion techniques in memes," in *Proc. 18th Int. Workshop Semantic Eval. (SemEval)*, Mexico City, Mexico, A. K. Ojha, A. S. Doğruöz, H. T. Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds., 2024, pp. 53–59.

[110] Y. Shu, H. Gu, P. Zhang, H. Zhang, T. Lu, D. Li, and N. Gu, "RAH! RecSys–assistant–human: A human-centered recommendation framework with LLM agents," *IEEE Trans. Computat. Social Syst.*, vol. 110, no. 5, pp. 6759–770, Jan. 2023.

[111] D. Chin, Y. Wang, and G. Xia, "Human-centered LLM-agent user interface: A position paper," 2024, *arXiv:2405.13050*.

[112] M. Luo, X. Xu, Y. Liu, P. Pasupat, and M. Kazemi, "In-context learning with retrieved demonstrations for language models: A survey," 2024, *arXiv:2401.11624*.



MARCO SIINO received the bachelor's and master's degrees (cum laude) in computer engineering from the University of Palermo, and the Ph.D. degree in information and communication technologies from the University of Palermo, in 2023. He is a freelance full stack Developer. He is currently an Assistant Professor with the University of Catania, where he teaches courses on network intelligence and information theory. His main interests include machine learning, deep learning, natural language processing, and recommender systems. His work involves social-networks-related tasks (e.g., sentiment analysis, hate speech detection, and fake news detection). He is an IEEE IMS Member.

MARIANA FALCO received the Ph.D. degree in engineering from Universidad Austral, in 2022, with a doctoral fellowship. She has been actively involved in research projects since 2013. She has over nine years of experience as a Lecturer in software engineering and project management. She has also managed software development projects as a Project Manager, specializing in agile methodologies for more than six years. In the past years, she has conducted audits on processes and the application of agile methodologies at scale in large companies. She has served as a reviewer for academic journals for the past three years, regarding software engineering and software quality.

DANIELE CROCE received the double M.Sc. degree in networking engineering from the Politecnico di Torino and EURECOM Institute, Sophia Antipolis, France, in 2006, the Research Master Diploma degree (ex DEA) in networking and distributed systems from the Université de Nice-Sophia Antipolis (UNSA), Nice, France, in 2006, and the joint Ph.D. degree from the Politecnico di Torino, Turin, Italy, and UNSA, in 2010. He is currently an Assistant Professor with the University of Palermo, Palermo, Italy. He has long experience of research collaborations, in several European and national research projects, on wireless networks, the Internet of Things, high-quality TV streaming, smart grid communications, and smart cities. He also worked on assistive technologies for visually impaired people and with the Arianna Project. He was the co-founder of the start-up company In.sight s.r.l., spin-off of Palermo University.

PAOLO ROSSO is a Full Professor of computer science with the Universitat Politècnica de València (UPV), Spain. He is a member of the Pattern Recognition and Human Language Technology (PRHLT) Research Center and the Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI). He has published more than 400 peer-reviewed papers in conferences and journals. He is among the best computer science scientists in Spain (https://research.com/scientists-rankings/computer-science/es). He is the PI of several related research projects, such as FairTransNLP-Stereotypes: Fairness and Transparency for equitable NLP applications in social media–Identifying stereotypes and prejudices and developing equitable systems (Grant PID2021-124361OB-C31), FAKEnHATE-PdC: FAKE news and HATE speech (Grant PDC2022-133118-I00), and XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (Grant PLEC2021-007681), funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTRPI. His current research interests include detection of harmful information in social media, both fake news and hate speech. In 2022, he received the UPV Research Award for Excellent Publication in Engineering and Technology on misogyny identification.