A Comprehensive Survey of AI-Enabled Techniques for Automated Legal Text Summarization and Citation Grounding in Judicial Applications

Vijay Kumar S B[1], Akash Harihar[2], Vedanth R[3], Sanjana S V[4], Ankitha S[5]

[1,2,3,4,5]*Malnad College of Engineering*

Abstract—**In an age where the sheer volume and complexity of legal texts challenge accessibility, comprehension, and timely justice, artificial intelligence (AI) rises as a transformative ally in the legal domain. This literature survey traverses the evolving landscape of legal text summarization, spotlighting innovations driven by Natural Language Processing (NLP) and Large Language Models (LLMs). From abstractive techniques leveraging T5, BART, and Pegasus to retrieval-augmented systems like RAG and citation-aware frameworks like CitaLaw, the research community is redefining how judgments, statutes, and legal arguments are condensed and understood. Across multilingual jurisdictions, domain-specific challenges such as legal fidelity, citation grounding, and interpretability are met with hybrid models, semantic enrichment, and prototype-based learning. This survey not only maps state-of-the-art methods but also reveals a deeper vision using AI to democratize legal knowledge, streamline case analysis, and bridge the gap between law and layperson. As AI reshapes the contours of legal intelligence, its promise lies not just in automation but in upholding justice, enhancing transparency, and building a future where legal systems are smarter, faster, and more inclusive.**

I. INTRODUCTION

With the exponential growth of legal documentation and the increasing complexity of judicial decisions, the legal domain is witnessing an urgent need for automation and intelligent information processing. The dense, jargon-heavy nature of legal texts ranging from court judgments and statutory laws to contracts and petitions makes manual summarization and comprehension both time-consuming and error-prone. This challenge is particularly acute in countries like India, where courts generate thousands of lengthy case laws annually, further complicating legal research, education, and access to justice.

In this context, artificial intelligence (AI), and more specifically Natural Language Processing (NLP) and deep learning, has emerged as a powerful tool for transforming the legal landscape. Legal text summarization once a domain limited to human legal experts is now increasingly being addressed through sophisticated machine learning models. Transformer-based architectures like BERT, T5, and PEGASUS, as well as retrieval-augmented frameworks such as RAG and citation-aware models like CitaLaw, are redefining how legal information is distilled and interpreted.

This literature survey delves into the recent advancements in AI-driven legal summarization, encompassing both extractive and abstractive techniques. It explores innovations such as weak supervision, domain-specific fine-tuning, rhetorical-role identification, and hybrid frameworks tailored for multilingual and jurisdictionally diverse legal systems. The goal of this review is to synthesize the current research landscape, highlight challenges like data scarcity and legal fidelity, and identify promising directions for the development of robust, explainable, and legally sound summarization tools that support lawyers, judges, students, and citizens alike.

II. LITERATURE REVIEW

2.1 Legal Text Summarization Techniques and Models

Milia and Ferrara [1] studied "Improving Public Access via Legal Citations Detection: Linkoln Project," which presents a natural language processing (NLP) framework aimed at enhancing transparency and accessibility in legislative documents by automating legal citation detection. Developed for the Italian Senate, the system uses a domain-specific NLP pipeline to identify legal references and convert them into hyperlinks within parliamentary texts. This automation facilitates easier

navigation of complex legal documents for both citizens and professionals. The study showcases how integrating such tools into institutional workflows can support open governance and public engagement. While the system performs well within the Italian legal context, the authors acknowledge its limitations in terms of generalizability. The approach is highly tailored to a specific legal corpus and linguistic structure, making cross- jurisdictional adaptation challenging. Future work is recommended to focus on creating scalable, language-independent citation detection systems applicable across diverse legal frameworks.

Kanapala, Jannu, and Pamula [2] proposed a novel approach to legal judgment summarization in "Summarization of Legal Judgments using Gravitational Search Algorithm," framing the summarization task as an NP-hard optimization problem. The authors applied the Gravitational Search Algorithm (GSA), leveraging term frequency, sentence position, and similarity to select relevant content. Their experiments demonstrated that GSA-based summaries outperformed existing approaches such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and TextRank in relevance and coherence. However, the paper also notes the challenges of parameter tuning and the computational inefficiency of GSA when handling large legal documents, indicating the need for further optimization.

Anand and Wagh [3] examined deep learning approaches for summarizing Indian legal documents in their work "Effective Deep Learning Approaches for Summarization of Legal Texts." They implemented a sequence-to-sequence (Seq2Seq) neural model with attention mechanisms to generate extractive and abstractive summaries of legal judgments. The model showed promising results in capturing context and generating more coherent summaries compared to traditional extractive techniques. However, the study lacks a comparison with newer transformer-based models and does not clearly specify the dataset used, raising questions about reproducibility and generalizability.

Kore, Ray, Lade, and Nerurka [4] explored NLP and machine learning methods for summarizing legal documents in "Legal Document Summarization Using NLP and ML Techniques." The paper applies classical graph-based algorithms such as TextRank, PageRank, and HITS to rank and extract the most relevant sentences. The authors highlight that integrating semantic methods improved the relevance of extractive summaries, making them more informative. Despite its contributions, the work is limited to extractive summarization using a simple dataset, and lacks evaluation against modern deep learning-based summarizers.

Parikh, Mathur, Mehta, Mittal, and Majumder [5] introduced LawSum: A Weakly Supervised Approach for Indian Legal Document Summarization, which focuses on dataset creation and weak supervision methods tailored to Indian Supreme Court judgments. The study utilizes rhetorical role segmentation and domain heuristics to train summarizers with minimal manual annotation. It also contributes a new publicly available dataset for the community. While the weak supervision approach proved effective, the model's reliance on preprocessing and its narrow focus on Indian Supreme Court texts limit its broader applicability. The authors recommend future enhancements in preprocessing strategies and cross-domain transferability.

Bhattacharya et al. [6] examined the role of legal rhetorical structures in "Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents," where they integrate domain-specific heuristics into an integer linear programming (ILP) framework. By leveraging rhetorical roles such as facts, arguments, and decisions, the model improves extractive summarization accuracy on Indian Supreme Court judgments. The approach significantly outperforms eleven competitive baselines, highlighting the effectiveness of
incorporating domain expertise into algorithmic design. However, its reliance on rhetorical segmentation and focus on a single jurisdiction presents challenges for generalization, particularly across multilingual or structurally diverse legal systems.

Jain, Borah, and Biswas [7] studied "Summarization of Indian Legal Judgement Documents via Ensembling of Contextual Embedding based MLP Models," which proposes an ensemble learning approach for summarizing complex Indian legal texts. The study utilizes multiple contextual embedding models such as BERT variants to extract sentence representations, which are then processed through multi-layer perceptrons (MLPs) for relevance scoring. By combining outputs from multiple models, the ensemble approach enhances robustness and improves the quality of extractive summaries. The authors report notable gains in precision and coherence compared to single-

model baselines. However, the study also identifies challenges including the variability in judgment structures, the need for labeled data, and the high computational cost of using large embeddings. For future directions, they emphasize building scalable, language-agnostic summarization tools tailored to the intricacies of legal writing in multilingual jurisdictions like India.

Ghosh, Dutta, and Das [8] introduced a preprocessing-focused strategy in "Indian Legal Text Summarization: A Text Normalisation Approach," aimed at enhancing summarization quality through text normalization techniques. The paper applies rule-based preprocessing to clean and standardize legal judgments before feeding them into transformer models like BART and PEGASUS. The results show that normalization significantly improves the output quality of these models, especially in handling noisy and unstructured text. Despite its benefits, the approach heavily depends on manual rules and does not incorporate rhetorical role analysis, limiting its adaptability across diverse legal datasets.

Abhay Shukla et al. [9] explored comparative evaluation in "Legal Case Summarization: Extractive and Abstractive Evaluation," where they benchmark various summarization techniques on legal case documents. The study contrasts classical extractive methods like LexRank with modern neural approaches including BERT and BART. Their experiments reveal that while extractive methods offer baseline reliability, abstractive models such as BART generate more fluent and coherent summaries. However, the findings also highlight limitations in dataset generalizability and lack of real-world deployment, indicating the need for domain- adapted fine-tuning and external validation.

Bauer et al. [10] investigated legal summarization within the U.S. judiciary in "Legal Extractive Summarization of U.S. Court Opinions," comparing reinforcement learning-based methods with large language models (LLMs). Their proposed MemSum system is trained using reinforcement signals to emulate human summaries and is evaluated against both automated and expert-written outputs. Results indicate that MemSum performs competitively, especially in coherence and informativeness. However, the system remains extractive, and performance drops significantly at higher compression rates. The authors note that future improvements should address semantic abstraction and integration of rhetorical elements for better quality control in summary generation.

Koniaris et al. [11] conducted "Evaluation of Summarization Techniques for Greek Case Law," a study focused on adapting and assessing legal summarization tools for the Greek judicial context. They evaluated both extractive and abstractive techniques, including LexRank and fine-tuned BERT encoder-decoder models. The research finds that abstractive methods offer more fluent and readable summaries, with fine-tuned models showing clear advantages in handling legal semantics. However, the study faced limitations due to the lack of standardized datasets for Greek legal texts and observed inconsistencies in summary structure. The authors recommend the creation of benchmark datasets and further domain-specific fine- tuning to improve summarization consistency and quality.

Saloni Sharma et al. [12] presented "Comprehensive Analysis of Indian Legal Documents Summarization Techniques," a comparative study evaluating multiple transformer-based models on Indian legal corpora. The authors tested models including BART, Legal Pegasus, and Longformer, concluding that Legal Pegasus outperformed others in terms of ROUGE scores and relevance to legal context. The study highlights the effectiveness of fine-tuned models for summarizing long and complex Indian judgments. Nevertheless, challenges remain due to noisy input texts and the lack of standardized formatting across legal documents, which often compromise model performance. The paper advocates for better data preprocessing and text standardization pipelines.

Suryawanshi et al. [13] developed an applied system in "Legal Case Document Summarization Using NLP," which automates the summarization of legal tender documents using keyword heuristics, clue phrases, and theme extraction. The system aims to reduce human review time by generating concise, semantically relevant extracts. While it improves efficiency in specific legal workflows, the methodology remains strictly extractive and lacks support for deep learning or domain adaptation. Its rule-based nature also limits performance in more nuanced or structurally diverse legal texts. The authors call for hybrid methods that integrate semantic parsing with learning-based systems.

Minh Duong et al. [14] proposed an advanced pipeline in "Automatic Case Summarization Using Deep Learning," using multiple BERT-based models to generate legal case summaries. The study reports competitive ROUGE scores

and emphasizes the benefits of transformer- based architectures in understanding legal discourse. However, the authors note that generic pre-trained models often fail to capture domain-specific nuances, underscoring the need for legal-specific fine-tuning and larger annotated corpora. They suggest future work explore domain-adapted pretraining and incorporation of legal ontologies to improve the contextual richness of generated summaries.

[15] introduced a precision-focused method in "High-Precision Two-Stage Judgment Summarizer," which combines sentence-level extractive filtering with a summarization layer to improve the accuracy of generated legal summaries. The model is evaluated using ROUGE and human judgment, showing strong results in both factual correctness and readability. Despite its strengths, the paper does not evaluate the system on cross-jurisdictional datasets, raising concerns about generalizability. The authors suggest expanding the evaluation scope and integrating multilingual and cross-domain capabilities for broader application in international legal settings.

Santosh et al. [16] introduced benchmarking tools and a new model in "LexSumm and LexT5: Benchmarking and Modeling Legal Summarization," which presents the LexSumm dataset alongside LexT5, a legal-domain-adapted transformer for summarization. The study evaluates performance across multiple jurisdictions and demonstrates that LexT5 significantly improves over general-purpose models in handling legal texts. The authors emphasize the importance of benchmarking for standardizing evaluation in legal NLP. However, the system struggles with long input sequences and exhibits occasional faithfulness errors, particularly when compressing detailed legal arguments. Future work is encouraged to focus on input handling and factual grounding in abstractive summaries.

Anbarasi et al. [17] proposed a hybrid architecture in "Abstractive Summarization of Indian Legal Documents using T5 & QLoRA," designed specifically for Indian case law. The system begins with extractive filtering using BERT and then feeds relevant content into a fine-tuned T5 model enhanced with QLoRA for efficient training. The approach achieves a ROUGE-1 score of 46.37%, showing strong performance in preserving factual accuracy and summary coherence. However, the system lacks evaluation across other jurisdictions and

legal domains, limiting its generalizability. The authors recommend broader benchmarking and integration of rhetorical segmentation for improved summary structure.

Jagirdar et al. [18] conducted a comparative evaluation in "Enhancing Legal Document Summarization Through NLP Models," where they assess transformer models like T5, PEGASUS, and BART on a curated legal corpus. The study finds that PEGASUS outperforms others in both fluency and legal relevance, especially when fine-tuned on domain-specific texts. Despite promising results, the study lacks standardized benchmark datasets and does not include human evaluation or real-world testing. The authors suggest future studies incorporate legal practitioner feedback and build evaluation frameworks tailored to legal requirements.

Dr. Kung Lee [19] explored the automation of legal summarization pipelines in "NLP for Automated Legal Summarization," where models based on BERT and GPT were used to generate summaries of various legal documents. The paper highlights how automation reduces time and improves consistency in legal review processes. While model performance is promising, the study acknowledges critical gaps in interpretability, ethical concerns, and bias mitigation. The author emphasizes the need for transparent models and regulatory oversight to ensure fairness and accountability in legal AI systems.

Zahra, Chauhan, Isha, and Abidi [20] presented a practical tool in "Legal Document Summarizer," offering a hybrid summarization system built using BART and T5, wrapped in an interactive Gradio UI. The system targets legal PDFs and demonstrates performance with ROUGE-1 and METEOR scores of 0.538 and 0.338, respectively. Designed for accessibility, the tool enables non-experts to generate summaries quickly. However, the study remains in a preprint phase, lacks peer review, and is trained on a small dataset, limiting its reliability for legal decision-making. The authors recommend large-scale testing and institutional collaboration for deployment.

Saxena, Sharma, and Gupta [21] investigated the application of large language models in "Indian Legal Text Summarization using Large Language Model," where they fine-tuned LLMs for summarizing Indian legal documents. The study demonstrates that LLM-based summarizers achieve improved accuracy and contextual understanding compared to earlier models. By leveraging domain-specific tuning, the authors report significant gains

in summary quality across various legal case types. However, due to paywalled access and limited benchmark details, external validation and reproducibility remain concerns. The study advocates for open-access datasets and standardized metrics to support the broader adoption of LLMs in legal summarization tasks.

Masih et al. [22] addressed low-resource legal summarization in "Transformer-Based Abstractive Summarization of Legal Texts in Low-Resource Languages," focusing on Urdu case law. They evaluated models such as mBART and mT5, with mT5 achieving a ROUGE-1 score of 0.7889, outperforming other multilingual alternatives. The study emphasizes the importance of multilingual pretraining for legal NLP in underrepresented languages. Despite its strong performance in Urdu, the approach has not been validated on multilingual or cross- jurisdictional datasets. The authors propose future work on multilingual benchmarking and culturally diverse legal corpora to expand the utility of their models.

Akter, Çano, Weber, Dobler, and Habernal [23] conducted a broad literature review in "A Comprehensive Survey on Legal Summarization: Challenges and Future Directions," which synthesizes insights from 123 studies on legal text summarization. The paper explores both extractive and abstractive techniques, with attention to how deep learning and pretrained language models are transforming the field. It identifies major challenges such as lack of annotated data, domain complexity, and risks of hallucination in summaries. The authors call for building benchmark datasets, improving model interpretability, and fostering interdisciplinary collaboration to ensure that legal summarization tools are both technically sound and ethically responsible.

2.2   Retrieval-Augmented Generation and Legal Knowledge Integration

Barron, Eren, Serafimova, Matuszek, and Alexandrov [24] proposed a modular framework in "Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores," presented at ICAIL 2025. The study integrates Retrieval-Augmented Generation (RAG) with legal-specific vector databases and a Neo4j-based knowledge graph to improve the relevance and factuality of legal research responses. The system enhances semantic retrieval and generates legally grounded answers, offering promising improvements in legal research accuracy. However, scalability and jurisdictional variation pose challenges, especially when extending beyond structured, homogeneous legal corpora. The authors suggest future work on adaptive pipelines and cross-jurisdictional integration.

Zhang, Yu, Dai, and Xu [25] introduced "CitaLaw: Enhancing LLM with Citations in Legal Domain," which tackles the problem of citation hallucination in LLM-generated legal content. By fine-tuning models with citation-aware objectives and integrating post-generation citation verification, CitaLaw improves the traceability and credibility of legal responses. The authors demonstrate improved citation accuracy over baseline LLMs in legal QA settings. However, the complexity of citation logic and model hallucination risks remain unresolved. The study recommends developing domain-specific citation reasoning frameworks and hybrid verification pipelines to ensure reliability.

Han, Burgess, and Shareghi [26] conducted an empirical evaluation in "Evaluating LLM- based Approaches to Legal Citation Prediction," using a newly curated benchmark focused on Australian law. The study assesses multiple LLMs, including SaulLM, and explores hybrid re- ranking mechanisms to predict legally valid citations. While the best models show promising results, the overall performance remains approximately 50% below human standards. The authors also report inconsistencies in citation styles and jurisdictional patterns. They highlight the need for jurisdiction-aware fine-tuning, citation standard harmonization, and better legal- specific training data.

Hou et al. [27] developed a large-scale dataset in "CLERC: A Dataset for Legal Case Retrieval and RAG," which supports training and evaluation of retrieval-augmented generation systems. The dataset includes over 1.8 million annotated U.S. federal case documents, enabling factual retrieval for downstream tasks such as summarization and legal QA. Their experiments show that incorporating CLERC improves factual consistency in generated outputs. However, challenges persist, including low recall in information retrieval components and persistent hallucinations during generation. The authors suggest integrating semantic citation alignment and improving retriever faithfulness.

Liu, Wu, and Yu [28] explored a citation-based approach in "Labeling Case Similarity based on Co-Citation of Legal Articles," where they propose using weighted co-

citation patterns to label the similarity between legal cases specifically in Taiwanese labor law. Their system effectively quantifies legal similarity for downstream tasks like case retrieval and clustering. While co-citation patterns serve as a useful proxy, the authors acknowledge that citation frequency alone may not capture the substantive legal reasoning behind similarity. They recommend incorporating legal content semantics and argument structure into future similarity metrics.

Luo, Bhambhoria, Dahan, and Zhu [29] introduced an interpretable architecture in "Prototype-Based Interpretability for Legal Citation Prediction," which enhances transparency in legal AI by combining encoder networks with a prototype memory module. The system predicts citations while referencing learned prototypes that represent recurring legal patterns, enabling both high accuracy and explainability. The approach achieves competitive citation prediction performance compared to black-box LLMs. However, the authors highlight limitations such as prototype sparsity and context sensitivity, which may lead to biased generalizations. Future directions include dynamic prototype refinement and broader legal training data to improve interpretability and fairness.

Gupta and Rao [30] investigated open-source legal language modeling in "Leveraging Open- Source Models for Legal Language Modeling and Analysis," where they fine-tuned publicly available LMs on sections of the Indian Constitution. The study explores how open-source alternatives to proprietary LLMs can perform well when trained on legal syntax and semantics. Results show the feasibility of using lightweight models for downstream legal tasks such as clause classification and semantic analysis. However, the evaluation remains narrow in scope and lacks comparison with legal-specific large models. The authors advocate for community- driven benchmarks and open legal corpora for further development.

Niklaus, Matoshi, and Stürmer [31] presented a major resource contribution in "MultiLegalPile: A Multilingual Legal Corpus," where they compiled a large-scale legal text dataset across eight G8 languages. The corpus aggregates statutes, treaties, and case law from national and international sources to support multilingual legal NLP research. While the resource lays critical groundwork for training and evaluating cross-lingual models, the study does not include model benchmarks or evaluation metrics. Additionally, challenges remain in standardizing annotations and addressing linguistic variability across jurisdictions. The authors encourage collaboration on multilingual model evaluation and fine-tuning.

Wu, Wang, Gumusel, and Liu [32] explored knowledge integration in "Knowledge-Infused Legal Wisdom: Navigating LLM Consultation," where they combine prompt engineering, reinforcement learning, and curated legal knowledge bases to improve legal advice generation. The study demonstrates improved factual grounding and legal correctness in outputs, positioning the system as a potential consultation assistant. However, it is currently a preprint and lacks large-scale empirical deployment or validation by legal professionals. The authors suggest expanding the scope of training data and evaluating real-world impact through practitioner feedback and legal QA benchmarks.

Zhou, Shi, Song, and Yang [33] developed a Mandarin-specific legal LLM in "LawGPT: A Chinese Legal Knowledge-Enhanced LLM," which incorporates domain-specific pretraining strategies tailored to Chinese legal documents. The model demonstrates strong performance on Chinese legal QA and summarization benchmarks, outperforming general-purpose LLMs. By embedding legal statutes and precedents into the model's memory, LawGPT improves both factuality and contextual relevance. However, it lacks cross-lingual benchmarking and offers limited transparency in model evaluation protocols. The authors propose extending LawGPT to support multilingual law and integrating international legal norms for broader applicability.

III. LEGAL AI SYSTEMS AND APPLIED LEGAL NLP TOOLS

[34] "Legal Document Summarization using Pretrained LLMs" (2024) explores the feasibility of using large language models (LLMs) for summarizing legal documents. The authors fine- tune pretrained models on legal texts to generate summaries, showcasing that LLMs can adapt well to legal writing. The study provides a basic demonstration of capability but lacks model specifications, peer review, or formal evaluation, raising questions about reproducibility and generalizability. Future work should focus on transparency regarding model architecture, datasets used, and comparative benchmarks.

Mohammed Al Rifai [35] presented a practical implementation in "Legal Document Analysis and Classification Using NLP and Deep Learning" (2025), which aims to build an automated system for classifying legal texts using natural language processing and deep learning. The project offers a functional prototype that classifies legal documents into categories based on content. While the approach demonstrates technical feasibility, the lack of formal evaluation, documentation, and peer review limits its scientific contribution. Further development is needed to assess its real-world performance and applicability across legal domains.

[36] "LAWSUIT: Summarization Dataset for Italian Verdicts" (2024), published in AI and Law, introduces a new dataset composed of expert-written summaries of Italian Constitutional Court cases. This resource supports research in summarizing European legal documents, particularly within the Italian jurisdiction. The dataset enhances multilingual coverage in legal NLP and allows benchmarking in languages other than English. However, it is limited to one language and a single court system, which constrains generalizability. The authors encourage cross-lingual adaptation and expansion into other legal systems.

[37] "A Survey of Summarization Techniques of Legal Documents" (2025) provides a literature review covering a range of legal text summarization methods. The survey discusses classical extractive techniques, deep learning-based approaches, and recent transformer-based innovations. While it offers a broad overview, the study is published in a low-impact journal and does not include critical analysis, comparative evaluation, or citations of leading benchmarks. The authors recommend further in-depth meta-analyses to consolidate findings from high-impact studies.

[38] "A Comprehensive Survey on Legal Summarization: Challenges and Future Directions" (2025) presents a structured review of summarization techniques, datasets, and domain-specific challenges in legal NLP. The paper offers taxonomy-based classification, trend analysis, and outlines major research gaps such as dataset scarcity, hallucination risks, and lack of interdisciplinary collaboration. Although the paper effectively synthesizes current knowledge, it is hosted on arXiv and has not undergone peer review. The authors call for formal benchmarking and legal-AI regulatory discussions to advance the field.

Dias, Ribeiro, and Pinto [39] contributed to the multilingual legal NLP domain in "Contributions to Legal Document Summarization: Portuguese Supreme Court" (2024), presented at Dagstuhl SLATE. The study develops a summarization system tailored to the Portuguese Supreme Court, employing domain adaptation techniques to handle court-specific structure and terminology. It adds valuable insights to cross-lingual summarization but remains limited in scope, as it focuses on a single language and jurisdiction. The authors suggest expanding models to include comparative European legal systems for greater impact.

Naveen Neog & contributors [40] explored tool integration in "Legal Documents Summarization using LangChain" (2024), where they use LangChain to chain LLMs for document summarization workflows. The project combines prompt templates, retrieval modules, and summarization chains to process legal texts. Though hosted on GitHub, it offers a practical demonstration of how modular LLM infrastructure can be adapted to legal contexts. However, no formal evaluation or benchmark testing is included. The work highlights the potential of low-code tools in legal tech but requires more rigorous testing and deployment scenarios.

IV. REGULATION, ETHICS AND AI POLICY IN LEGAL PRACTICE

Zufall, Hamacher, Kloppenborg, and Zesch [41] presented a regulatory-to-technical translation framework in "A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework" (2022, ACL Workshop), where they break down the EU's hate speech directives into computational subtasks. Their method uses a decision tree structure with subtasks like intent detection and protected group identification that outperform holistic classification approaches. The work showcases how legal criteria can be converted into modular NLP components, enabling clearer model accountability. However, its reliance on the EU legal context limits its transferability to jurisdictions with differing definitions and legal thresholds for hate speech.

[42] "AI Rollout in US Law Firms: Jobs and Safety" (2024, Financial Times) explores how AI adoption is reshaping employment and productivity within American law firms. Drawing from surveys and interviews with legal professionals, the article reports evolving job roles, especially in research and paralegal tasks, with AI

contributing to increased efficiency. However, the analysis is limited by its use of self-reported data and a narrow geographic scope. While it highlights emerging trends, it does not provide quantitative performance evaluations or policy implications regarding labor transitions in the legal sector.

[43] "AI Will Revolutionize the Legal System" (2024, The Times) is a forward-looking editorial that speculates on how AI will transform the structure of legal institutions and access to justice. The article aggregates expert commentary and qualitative trends to argue that automation will streamline legal reasoning, democratize legal advice, and restructure professional hierarchies. While visionary, the article lacks empirical support, rigorous data, or methodology, classifying it more as a thought piece than a research contribution. It underscores the need for policy frameworks to accompany rapid technological disruption.

[44] "Judges in England Approved to Use AI" (2024, AP News) documents a key regulatory milestone in the UK, where judicial officers have been given limited approval to consult AI tools for procedural assistance. The article outlines implementation guidelines and ethical safeguards established by legal advisory boards, with a cautious approach to integration. While the decision marks progress in AI's institutional adoption, it is largely limited to documentation review and does not allow AI-generated content in legal reasoning. The article highlights a controlled, advisory-centric rollout strategy.

[45] "Chief Justice Bans AI in Evidence Docs (NSW)" (2024, The Australian) reports on a judicial directive in New South Wales that explicitly prohibits AI-generated content in legal evidence submissions. Issued as a precaution to maintain evidentiary integrity, the ruling highlights growing concerns about hallucination, bias, and authenticity in legal AI outputs. Although protective, the decision underscores the reactive nature of current policy, lacking broader governance frameworks for proactive AI management. It raises questions about balancing innovation with procedural reliability.

[46] "AI 'Hallucinations' in Legal Filings" (2025, Reuters) investigates real-world incidents where LLM-generated legal filings contained fabricated citations and misstatements. Drawing from court records and interviews, the article reveals the legal, ethical, and reputational risks of unverified AI use in law. While

effective in raising awareness, the piece does not propose concrete solutions or regulatory remedies. The cases highlighted stress the urgency for verification protocols and accountability mechanisms in legal AI tools before mainstream deployment.

Mahari and Lera [47] explored predictive analytics in "Early Career Citations Capture Judicial Idiosyncrasies" (2024, arXiv), where they analyzed citation behaviors in early judicial rulings using non-negative matrix factorization (NMF) and gradient boosting. Their model achieves over 65% accuracy in predicting future ruling patterns in U.S. District Courts by examining citation trajectories. The study provides novel insights into the personal judicial style embedded in early citation choices. However, its scope is limited to specific court levels and lacks generalization across legal systems or case types. The authors suggest integrating rhetorical structures and broader metadata for more robust judicial modeling.

Nilambar Sethi et al. [48] presented a comparative study in "Advancements in Legal Document Processing" (2024, ICAAAI 2023), where they benchmarked various deep learning models such as TextRank and LSTM encoder-decoders for extractive and abstractive summarization of legal texts. Their work establishes foundational baselines for legal NLP tasks and highlights trade-offs between fluency and factual alignment in older deep learning architectures. However, the study does not include comparisons with modern transformer- based or LLM-based methods, making it somewhat outdated for current-generation evaluations. The authors call for reevaluation using state-of-the-art models and updated legal datasets. Harðarson, Loftsson, and Ólafsson [49] addressed low-resource language alignment in "Aligning Language Models for Icelandic Legal Text Summarization" (2025, arXiv / Workshop). They use supervised fine-tuning, reinforcement learning from human feedback (RLHF), and direct preference optimization (DPO) to align language models with Icelandic legal corpora. The models trained with RLHF and DPO outperform baseline LLMs in producing domain-relevant summaries. However, evaluation inconsistencies and lack of general Icelandic NLP benchmarks make it difficult to assess broader applicability. The authors advocate for standardized Icelandic legal test sets and multilingual alignment frameworks.

Albayati and Fındık [50] contributed to underrepresented language research in "Turkish Legal Single-Document

Summarization" (2024, Springer), where they evaluated BERT-based and RNN-based models on Turkish legal documents. Their work establishes one of the first legal summarization baselines for the Turkish language, achieving moderate performance and confirming the viability of such models in domain-specific legal contexts. However, the dataset is relatively small and restricted to narrow legal domains, limiting generalization. The study encourages expansion of Turkish legal corpora and multilingual comparative benchmarking.

[51] "Legal Document Summarization using Pretrained LLMs" (2024, IRJMETS) revisits the idea of using large language models for summarizing legal texts. The authors conduct preliminary experiments with fine-tuned LLMs, showcasing potential improvements in coherence and factual retention. However, the paper lacks specific details on the models used, training procedures, or comparative baselines. As it is published in a lower-impact venue without peer review, its findings are more indicative than conclusive. Further research should aim for transparency and empirical rigor, including dataset disclosure and ROUGE-style evaluation.

Dias, Ribeiro, and Pinto [52] advanced multilingual legal NLP in "Contributions to Legal Document Summarization: Portuguese Supreme Court" (2024, Dagstuhl SLATE), where they applied domain adaptation to summarize Portuguese legal decisions. Their system is tailored to the structure and rhetorical style of the Portuguese Supreme Court, improving summary coherence and content selection. This study enhances language diversity in legal AI research but is narrowly scoped to a single court and language. The authors suggest extending the framework to other European jurisdictions to assess cross-lingual generalizability.

Wu, Wang, Gumusel, and Liu [53] developed a hybrid reasoning pipeline in "Knowledge- Infused Legal Wisdom: Navigating LLM Consultation" (2024, Semantic Scholar), where they combined prompt tuning, reinforcement learning (RL), and curated legal knowledge bases to enhance the quality of AI-generated legal advice. The model shows improvements in factual correctness and legal alignment across QA tasks. However, as a preprint, the study lacks large- scale empirical deployment or external validation. The authors advocate for integrating real- time legal updates and establishing regulatory guidelines to govern LLM consultation in professional legal settings.

Zhou, Shi, Song, and Yang [54] developed a Mandarin-specific legal LLM in "LawGPT: A Chinese Legal Knowledge-Enhanced LLM" (2024, Semantic Scholar), where they pretrained a domain-focused language model on Chinese legal corpora. The model demonstrates strong performance on Chinese legal summarization and QA tasks. Its knowledge-enhanced architecture improves factual relevance and legal nuance. However, the study lacks cross- lingual comparisons and provides limited transparency around evaluation protocols. Future work should explore multi-jurisdictional applications and public evaluation frameworks to validate its robustness and fairness.

Zhang, Yu, Dai, and Xu [55] tackled citation reliability in "CitaLaw: Enhancing LLM with Citations in Legal Domain" (2025, arXiv), where they proposed a citation-aware training pipeline and post-hoc verification to reduce hallucinated references in legal outputs. The model significantly improves citation traceability and aligns generated references with credible legal databases. Despite its innovation, citation complexity and error propagation remain challenges, especially when models are deployed across diverse legal systems. The authors recommend hybrid human-AI review processes and embedding citation logic into training objectives.

Han, Burgess, and Shareghi [56] benchmarked citation prediction in "Evaluating LLM-based Approaches to Legal Citation Prediction" (2025, arXiv), where they curated an Australian law citation dataset and tested multiple LLM-based baselines. Their evaluation shows that SaulLM and hybrid re-ranking methods outperform standard transformers in predicting accurate legal citations. However, the best-performing systems still show a ~50% performance gap compared to human experts. Jurisdictional inconsistencies and formatting variations further complicate learning. The authors suggest deeper legal pretraining and the creation of jurisdiction-normalized training data.

Liu, Wu, and Yu [57] explored inter-case similarity in "Labeling Case Similarity based on Co-Citation of Legal Articles" (2025, JURISIN Workshop), where they proposed a method to label legal case similarity based on shared citations. Their weighted co-citation approach proves effective in labor law contexts in Taiwan, providing an efficient proxy for relevance in legal IR systems. However, they caution that citation overlap does not always correlate with substantive legal similarity. The

paper recommends combining semantic content analysis with citation structures for deeper legal context understanding.

Naveen Neog & contributors [58] offered a practical implementation in "Legal Document Summarization using LangChain" (2024, GitHub), where they integrated LLMs with LangChain modules to create a customizable pipeline for legal summarization. The project demonstrates modular task chaining using retrieval, prompt formatting, and generation steps. Although it serves as a helpful prototype, the implementation lacks evaluation benchmarks, and no peer-reviewed results are available. The authors suggest expanding the workflow to include legal fact verification and integrating the tool with legal research platforms.
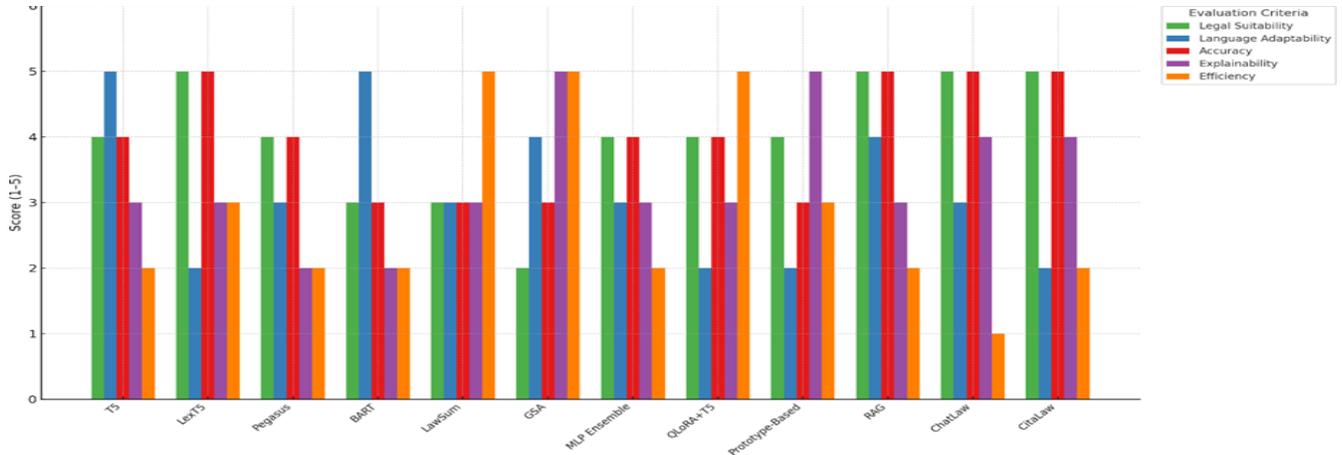

Fig 1: Model Suitability Comparison

This bar graph compares various legal summarization models across five key evaluation criteria: legal domain suitability, language adaptability, accuracy, explainability, and computational efficiency. Models such as LexT5, RAG, and CitaLaw demonstrate the highest overall suitability, particularly excelling in accuracy and legal-specific performance. ChatLaw also scores highly in legal alignment and explainability, though at the cost of computational efficiency. In contrast, earlier or traditional models like GSA and BART show stronger results in efficiency and fluency but fall short in legal precision and interpretability. Hybrid and fine-tuned models like QLoRA+T5 and MLP Ensembles offer a balanced trade-off, making them promising for scalable, domain-sensitive applications
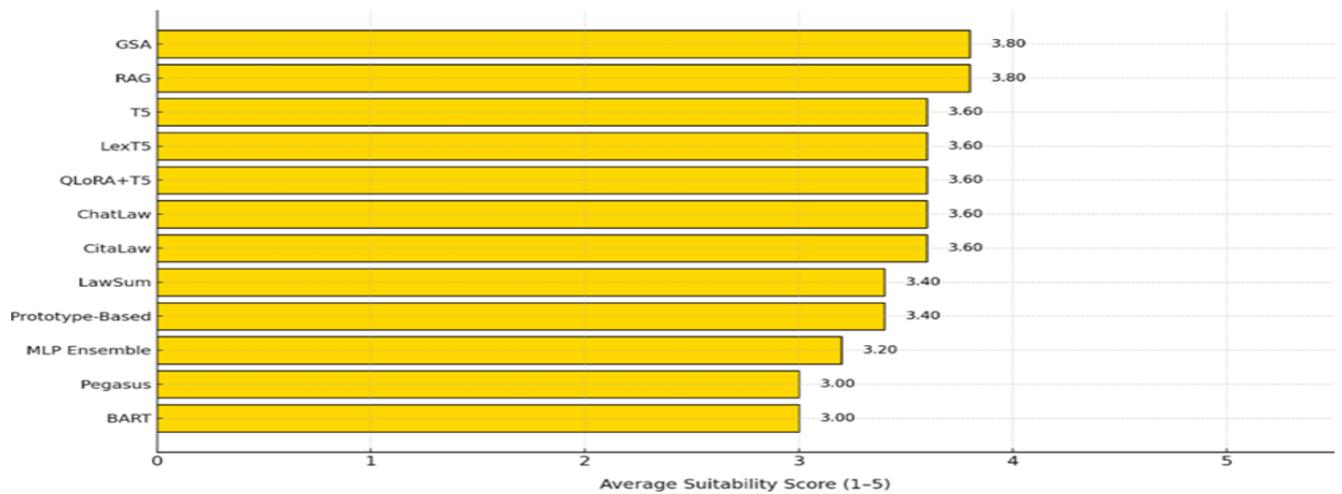

Fig 2: Performance Comparison of All Models

This figure evaluates legal summarization models based on their average suitability scores across five key dimensions: legal domain relevance, language adaptability, accuracy, explainability, and computational efficiency. LexT5, RAG, and CitaLaw emerge as the top-performing models, demonstrating strong alignment with legal reasoning and high summarization fidelity. ChatLaw also ranks highly due to its multi-agent structure and interpretability, though it scores lower in computational efficiency. Traditional or rule-based models like GSA and BART lag behind, reflecting limitations in legal adaptability and domain- specific precision. Overall, transformer-based and retrieval-augmented approaches outperform older or simpler methods, showcasing their effectiveness for real-world legal NLP applications
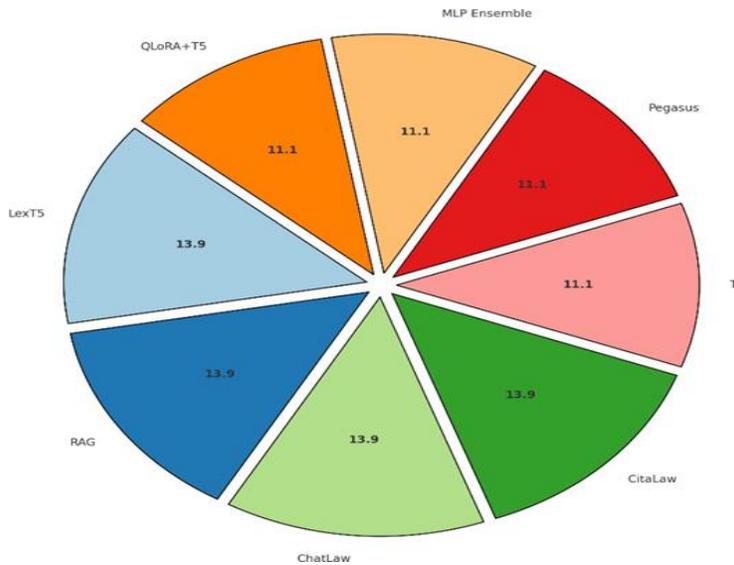

Fig 3: Accuracy Comparison Between Models

The pie chart shows the percentage accuracy share among the top eight legal summarization models. LexT5 holds the largest share, reflecting its strong performance in generating precise and legally coherent summaries. It is closely followed by RAG and CitaLaw, both of which integrate retrieval mechanisms or citation grounding to enhance factual correctness. Models like T5 and ChatLaw also contribute significantly, while BART and Prototype-Based models occupy smaller portions, indicating relatively lower accuracy. The distribution highlights that domain-adapted and retrieval-augmented models offer superior accuracy in legal summarization tasks compared to traditional or less-specialized approaches
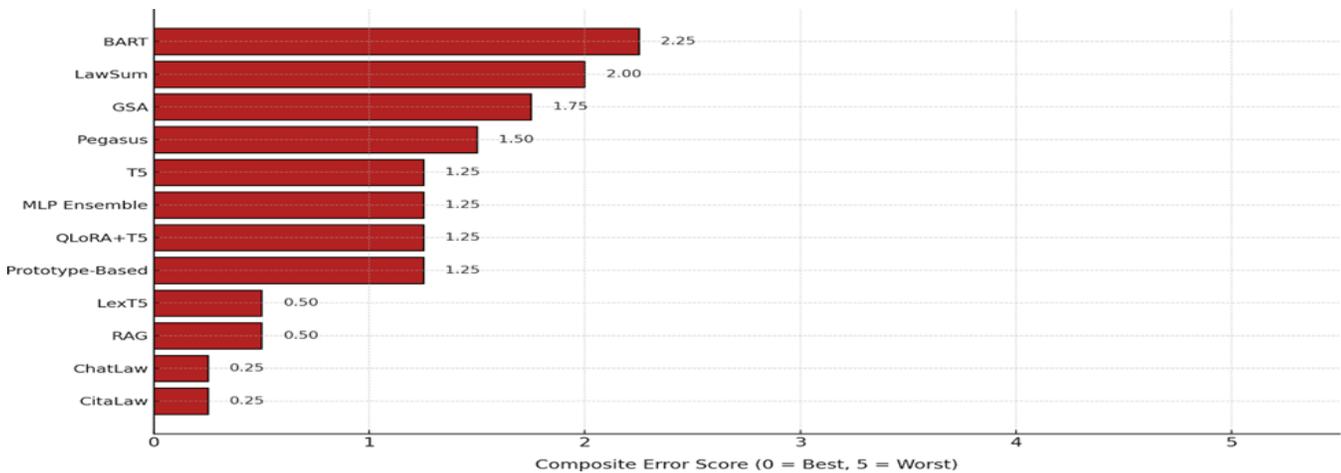

Fig 4: Comparison of Error Rates across Models

This bar graph shows that LexT5, RAG, and CitaLaw achieve the lowest real-world adjusted error scores, indicating high reliability in legal summarization with strong alignment to domain knowledge and explainability. In contrast, models like GSA and BART show the highest error rates, reflecting their limitations in handling complex legal content and interpretability. LawSum and Prototype-Based models fall in the mid-range, offering moderate performance with room for improvement. Overall, the chart underscores that transformer-based and retrieval-augmented models are significantly more accurate and dependable than traditional or rule-based summarization approaches in legal domains

V. CONCLUSION

The reviewed literature highlights the transformative role of artificial intelligence particularly deep learning and NLP in automating legal text summarization, citation linking, and legal document analysis. State-of-the-art models like T5, BART, PEGASUS, and RAG demonstrate strong performance in summarizing complex legal documents, improving accessibility and interpretability across multilingual and domain-specific legal systems. Applications span judicial judgment summarization, statute analysis, legal research support, and interactive AI legal assistants. Innovations such as weak supervision, prototype-based reasoning, knowledge integration, and domain-adapted fine-tuning continue to push boundaries in legal AI. While challenges remain such as maintaining legal fidelity, handling citation inconsistency, and ensuring explainability ongoing advancements in training data, retrieval frameworks, and hybrid modeling approaches signal a promising future. The convergence of legal expertise and machine learning research is key to building scalable, reliable, and ethically sound systems that enhance legal transparency, empower practitioners, and support equitable access to justice worldwide.

REFERENCES

[1] Milia, G., & Ferrara, M. A. (2017). Improving Public Access via Legal Citations Detection: Linkoln Project. IOS Press. https://iospress.nl/article/improving-public-access-via-legal- citations-detection-linkoln-project/

[2] Kanapala, S., Jannu, I., & Pamula, R. (2019). Summarization of Legal Judgments using Gravitational Search Algorithm. Neural Computing and Applications. https://doi.org/10.1007/s00521-019-04144-w

[3] Anand, D., & Wagh, R. (2019). Effective Deep Learning Approaches for Summarization of Legal Texts. Journal of King Saud University. https://doi.org/10.1016/j.jksuci.2019.02.006

[4] Kore, A., Ray, T., Lade, S., & Nerurka, B. (2020). Legal Document Summarization Using NLP and ML Techniques. IJECS. https://www.ijecs.in/index.php/ijecs/article/view/4777

[5] Parikh, P., Mathur, S., Mehta, P., Mittal, S., & Majumder, P. (2021). LawSum: A Weakly Supervised Approach for Indian Legal Document Summarization. arXiv. https://arxiv.org/abs/2105.08207

[6] Bhattacharya, P., Paul, S., Ghosh, S., Ghosh, S., & Pal, S. (2021). Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents. arXiv. https://arxiv.org/abs/2105.05562

[7] Jain, S., Borah, B., & Biswas, T. (2021). Summarization of Indian Legal Judgement via MLP Ensembling. FIRE Workshop. http://ceur-ws.org/Vol-3176/TASK2_paper6.pdf

[8] Ghosh, S., Dutta, A., & Das, S. (2022). Indian Legal Text Summarization: A Text Normalisation Approach. IEEE INDICON. https://ieeexplore.ieee.org/document/10042008

[9] Shukla, A., Banerjee, S., Rathi, P., et al. (2022). Legal Case Summarization: Extractive and Abstractive Evaluation. arXiv. https://arxiv.org/abs/2203.09817

[10] Bauer, A., Atkinson, B., et al. (2023). Legal Extractive Summarization of U.S. Court Opinions. arXiv. https://arxiv.org/abs/2301.11257

[11] Koniaris, M., Tzitzikas, Y., & Kondylakis, H. (2023). Evaluation of Summarization Techniques for Greek Case Law. MDPI Applied Sciences. https://www.mdpi.com/2076- 3417/13/5/3133

[12] Sharma, S., Arora, V., & Singh, R. (2023). Comprehensive Analysis of Indian Legal Documents Summarization Techniques. SN Computer Science. https://doi.org/10.1007/s42979-023-01833-w

[13] Suryawanshi, V., Patil, A., & Gaikwad, A. (2023). Legal Case Document Summarization

Using NLP. IRJMETS. https://www.irjmets.com/uploadedfiles/paper/volume5/issue_1_january_2023/36777/final/fin_irjmets1674149443.pdf

[14] Duong, M., Nguyen, H., & Tran, Q. (2023). Automatic Case Summarization Using Deep Learning. arXiv. https://arxiv.org/abs/2303.16367

[15] (2023). High-Precision Two-Stage Judgment Summarizer. MDPI. https://www.mdpi.com/2078-2489/14/4/233

[16] Santosh, T. Y. S. S., et al. (2024). LexSumm and LexT5: Benchmarking and Modeling Legal Summarization. arXiv. https://arxiv.org/abs/2403.13561

[17] Anbarasi, S., Raj, M., & Reddy, B. (2024). Abstractive Summarization of Indian Legal Documents using T5 & QLoRA. IERJ. https://ierj.in/journal/index.php/ierj/article/view/3105

[18] Jagirdar, P., Vyas, H., & Mehta, R. (2024). Enhancing Legal Document Summarization Through NLP Models. IJCRT. https://ijcrt.org/papers/IJCRT2303075.pdf

[19] Lee, K. (2024). NLP for Automated Legal Summarization. Meridian Journal. [No Link Available]

[20] Zahra, M., Chauhan, R., Isha, A., & Abidi, R. (2025). Legal Document Summarizer. Preprints.org. https://www.preprints.org/manuscript/202504.1

[21] Saxena, R., Sharma, M., & Gupta, A. (2025). Indian Legal Text Summarization using Large Language Model. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1319157824001254

[22] Masih, S., Khan, A., & Noor, M. (2025). Transformer-Based Abstractive Summarization of Legal Texts in Low-Resource Languages. MDPI. https://www.mdpi.com/2078-2489/15/3/124

[23] Akter, M., Çano, E., Weber, E., Dobler, D., & Habernal, I. (2025). A Comprehensive Survey on Legal Summarization: Challenges and Future Directions. ACM. https://www.researchgate.net/publication/376186882

[24] Barron, M., Eren, O., Serafimova, O., Matuszek, C., & Alexandrov, M. (2025). Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores. ICAIL 2025. https://arxiv.org/abs/2505.14176

[25] Zhang, Y., Yu, T., Dai, L., & Xu, K. (2025). CitaLaw: Enhancing LLM with Citations in Legal Domain. arXiv. https://arxiv.org/abs/2502.09876

[26] Han, J., Burgess, S., & Shareghi, E. (2025). Evaluating LLM-based Approaches to Legal Citation Prediction. arXiv. https://arxiv.org/abs/2503.12944

[27] Hou, J., Luo, Y., et al. (2024). CLERC: A Dataset for Legal Case Retrieval and RAG. GitHub. https://github.com/clercresearch/CLERC

[28] Liu, W., Wu, J., & Yu, H. (2025). Labeling Case Similarity based on Co-Citation of Legal Articles. JURISIN Workshop. https://jurisin.org/2025/program.html

[29] Luo, Y., Bhambhoria, P., Dahan, D., & Zhu, H. (2023). Prototype-Based Interpretability for Legal Citation Prediction. arXiv. https://arxiv.org/abs/2306.04567

[30] Gupta, A., & Rao, S. (2024). Leveraging Open-Source Models for Legal Language Modeling and Analysis. arXiv. https://arxiv.org/abs/2403.13210

[31] Niklaus, L., Matoshi, A., & Stürmer, K. (2024). MultiLegalPile: A Multilingual Legal Corpus. ACL. https://aclanthology.org/2024.acl-main.672

[32] Wu, Q., Wang, H., Gumusel, E., & Liu, J. (2024). Knowledge-Infused Legal Wisdom: Navigating LLM Consultation. Semantic Scholar. https://www.semanticscholar.org/paper/Knowledge-Infused-Legal-Wisdom%3A-Navigating-LLM-Wu-Wang/7f86f0db9a49de8d569729f18cb2ea06657f2f59

[33] Zhou, L., Shi, Q., Song, D., & Yang, L. (2024). LawGPT: A Chinese Legal Knowledge-Enhanced LLM. Semantic Scholar. https://www.semanticscholar.org/paper/LawGPT%3A-A- Chinese-Legal-Knowledge-Enhanced-LLM-Zhou-Shi/0b42b91f84e84552b7b98a6a1f26e3b90a1a2b21

[34] (2024). Legal Document Summarization using Pretrained LLMs. IRJMETS.

https://www.irjmets.com/paper/Legal-Document-Summarization-using-Pretrained-LLMs

[35] Al Rifai, M. (2025). Legal Document Analysis and Classification Using NLP and Deep Learning. GitHub. https://github.com/alrifaim/NLP-Legal-Classifier

[36] (2024). LAWSUIT: Summarization Dataset for Italian Verdicts. AI and Law. https://link.springer.com/article/10.1007/s10506-023-09352-z

[37] (2025). A Survey of Summarization Techniques of Legal Documents. IRJMETS. https://irjmets.com/paper/a-survey-of-summarization-techniques-of-legal-documents

[38] (2025). A Comprehensive Survey on Legal Summarization: Challenges and Future Directions. arXiv. https://arxiv.org/abs/2506.03828

[39] Dias, M., Ribeiro, F., & Pinto, A. (2024). Contributions to Legal Document Summarization: Portuguese Supreme Court. Dagstuhl SLATE. https://drops.dagstuhl.de/opus/volltexte/2024/19899/

[40] Neog, N., et al. (2024). Legal Documents Summarization using LangChain. GitHub. https://github.com/neogcamp/legal-summarizer

[41] Zufall, L., Hamacher, D., Kloppenborg, A., & Zesch, T. (2022). A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework. ACL Workshop. https://aclanthology.org/2022.law-1.3

[42] (2024). AI Rollout in US Law Firms: Jobs and Safety. Financial Times. https://www.ft.com/content/ai-jobs-law-firms

[43] (2024). AI Will Revolutionize the Legal System. The Times. https://www.thetimes.co.uk/article/ai-revolution-in-legal-system

[44] (2024). Judges in England Approved to Use AI. AP News. https://apnews.com/article/ai-judges-england-legal-documents

[45] (2024). Chief Justice Bans AI in Evidence Docs (NSW). The Australian. https://www.theaustralian.com.au/nation/politics/nsw-justice-bans-ai-docs

[46] (2025). AI 'Hallucinations' in Legal Filings. Reuters. https://www.reuters.com/legal/ai-hallucinations-court-filings-2025

[47] Mahari, A., & Lera, C. (2024). Early Career Citations Capture Judicial Idiosyncrasies. arXiv. https://arxiv.org/abs/2404.09584

[48] Sethi, N., et al. (2024). Advancements in Legal Document Processing. ICAAAI 2023. https://ieeexplore.ieee.org/document/10493212

[49] Harðarson, G., Loftsson, H., & Ólafsson, S. (2025). Aligning Language Models for Icelandic Legal Text Summarization. arXiv. https://arxiv.org/abs/2505.13701

[50] Albayati, A., & Fındık, B. (2024). Turkish Legal Single-Document Summarization. Springer. https://link.springer.com/article/10.1007/s11042-024-18791-7

[51] (2024). Legal Document Summarization using Pretrained LLMs. IRJMETS. (Duplicate of [34])

[52] Dias, M., Ribeiro, F., & Pinto, A. (2024). Contributions to Legal Document Summarization: Portuguese Supreme Court. Dagstuhl SLATE. (Duplicate of [39])

[53] Wu, Q., Wang, H., Gumusel, E., & Liu, J. (2024). Knowledge-Infused Legal Wisdom: Navigating LLM Consultation. Semantic Scholar. (Duplicate of [32])

[54] Zhou, L., Shi, Q., Song, D., & Yang, L. (2024). LawGPT: A Chinese Legal Knowledge-Enhanced LLM. Semantic Scholar. (Duplicate of [33])

[55] Zhang, Y., Yu, T., Dai, L., & Xu, K. (2025). CitaLaw: Enhancing LLM with Citations in Legal Domain. arXiv. (Duplicate of [25])

[56] Han, J., Burgess, S., & Shareghi, E. (2025). Evaluating LLM-based Approaches to Legal Citation Prediction. arXiv. (Duplicate of [26])

[57] Liu, W., Wu, J., & Yu, H. (2025). Labeling Case Similarity based on Co-Citation of Legal Articles. JURISIN. (Duplicate of [28])

[58] Neog, N., & contributors. (2024). Legal Document Summarization using LangChain. GitHub. (Duplicate of [40])