Large Language Models in Legal Analysis

Jens Frankenreiter & Michael A. Livermore

A. Introduction

In the five years since our comprehensive review of computational methods in legal analysis (Frankenreiter & Livermore 2020), the field has undergone a revolutionary transformation driven by the emergence of large language models (LLMs). What began as a promising but limited set of natural language processing tools has evolved into sophisticated systems capable of analyzing and generating legal texts with remarkable fluency and accuracy. GPT-3.5, which was released in 2022, demonstrated that neural language models could perform legal tasks ranging from contract analysis to bar exam questions. Subsequent models like GPT-4, Claude, and specialized legal LLMs have pushed these capabilities even further, achieving performance levels that often rival trained legal professionals.

This technological leap represents more than just an incremental improvement in computational power – it constitutes a paradigm shift in how we approach the intersection of law and artificial intelligence. Where earlier computational methods required extensive preprocessing, feature engineering, and domain-specific architectures, modern LLMs can work directly with legal texts in their natural form, modeling context, capturing nuance, and replicating legal reasoning patterns that previously required human expertise to capture. In contrast to earlier methods, they also do not require the construction of specialized architectures for each new task. As a result, researchers can integrate these models into their work relatively easily, both streamlining and scaling existing lines of inquiry, and opening entirely new approaches to measuring legally relevant phenomena. These developments are also reflected in Christoph Engel's recent scholarship, which – in line with his longstanding openness to methodological innovation – has recently turned to the integration of large language models and other machine learning techniques into legal analysis. This includes a paper coauthored by all three of us (with additional collaborators) (Dominguez-Olmedo et al. 2024).

The implications extend beyond academic research. Legal practitioners are increasingly adopting LLM-powered tools for document review, legal research, contract drafting, and case prediction. Law schools are grappling with how to integrate these technologies into curricula and assessment. Courts and regulatory bodies are beginning to encounter LLM-generated content and must develop frameworks for its evaluation and admissibility.

This updated review builds on our earlier framework while addressing the unique characteristics and applications of LLMs in legal contexts. We focus on three primary application areas that have emerged as dominant in the LLM era: sophisticated data generation and feature extraction; advanced inference and prediction capabilities; and engineering-focused research that treats legal tasks as benchmarks for model development and optimization. We maintain the foundational distinction between "law as code" and "law as data" from our prior work but also examine how LLMs blur these traditional boundaries.

As we will demonstrate, LLMs represent both the culmination of decades of progress in computational legal analysis and the beginning of an entirely new chapter in the field's evolution.

B. From Traditional NLP to Large Language Models

The emergence of large language models represents a fundamental transformation in how computational methods can be applied to legal analysis. This shift extends and complicates the traditional distinction between "law as code" and "law as data" approaches (Frankenreiter & Livermore 2020; Livermore and Rockmore 2019).

Law as code conceives of legal rules as logical statements that can be formalized into executable algorithms – decision trees that mechanically apply legal rules to factual inputs. These systems have found practical success in domains like tax preparation software, where clear rules can be translated into deterministic computational processes. Law as data approaches, by contrast, treat legal texts as sources of information to be analyzed quantitatively, using techniques from natural language processing and machine learning to extract patterns and insights from large corpora of legal documents.

LLMs treat law as data: they are trained on vast collections of legal texts to learn statistical patterns in legal language and reasoning.

However, they differ from earlier approaches in at least two important respects. First, unlike earlier computational methods that required extensive preprocessing and feature engineering – transforming legal texts into structured numerical representations – LLMs can work directly with legal language in its natural form. Second, LLMs learn patterns in legal language and reasoning through exposure to massive datasets, not through predefined feature sets or models – often unsupervised – that could capture only limited structural or topical patterns in text. This reduces the cost of tasks like feature extraction, which previously required bespoke training datasets, and increases interpretability for tasks that once relied on sometimes opaque statistical groupings by producing outputs in natural language.

This paradigmatic shift has profound implications for legal text processing. Traditional bag-of-words approaches reduced legal documents to collections of isolated terms, losing crucial information about word order, context, and semantic relationships. Topic models, while able to automatically discover thematic patterns in legal corpora, remained fundamentally limited by their bag-of-words foundation. Supervised machine learning approaches required extensive feature engineering for each task – researchers had to specify which textual patterns or document characteristics to extract for the relevant prediction task. LLMs, by contrast, concentrate engineering effort in pre-training on massive text corpora, learning general representations that capture vast amounts of semantic and contextual information. These learned representations can then be applied across diverse legal tasks without task-specific feature engineering, enabling analysis of complex textual dependencies that would be difficult or impossible to specify manually.

Beyond these advantages within law as data, LLMs have also been proposed as a bridge to law as code approaches [Ash, same volume]. One promising direction involves integrating formalized representations of legal texts with LLM-based classifiers capable of addressing vague terms (*cf.* Livermore 2020). For example, while classic law-as-code projects like formalizing the British Nationality Act faced difficulties with vague terms such as good character, an LLM-based classifier could analyze natural language descriptions of individual cases and generate an output concerning whether they satisfy such standards. Similarly, Janatian et al. (2024) demonstrated how LLMs can automatically extract structured representations from legislation, using GPT-4 to create decision pathways, with 60 % of generated pathways rated as equivalent or better than manually created

ones. This suggests LLMs can ease the costly development of transparent, rule-based systems while preserving their explainability advantages.

This convergence suggests that LLMs can enable more fluid translation between different representational formats, capturing both the rule-like characteristics of law as well as broader social, cultural, linguistic, and political contexts [Ash, same volume]. For doctrinal scholars, LLMs can assist in synthesizing vast bodies of case law, identifying doctrinal inconsistencies, and tracing the evolution of legal concepts across jurisdictions and time periods at unprecedented scale. For empirical researchers, LLMs can facilitate large-scale studies that can simultaneously capture broad patterns in legal decision-making while preserving the contextual richness that traditional computational methods often sacrifice.

In line with our previous review, this article focuses primarily on the second line of work, while also covering related research that LLMs have opened up for empirical scholars, including benchmarking and engineering research oriented toward improving model performance on legal tasks. The scholars and projects highlighted in the following sections illustrate the breadth and transformative potential of this emerging research agenda. Due to space constraints, we exclude other relevant areas, such as socio-legal examinations of how LLMs are affecting legal institutions and law practice.

C. Data Generation and Feature Extraction

In our previous review, we documented how computational analysis often serves as a first step in empirical legal research, generating structured data from unstructured legal texts for use with traditional statistical techniques, including causal inference.[1] Large language models extend this line of work in important ways. First, they enable the extraction of features that depend on multiple stylistic and semantic dimensions, such as patterns in legal reasoning styles. Earlier approaches often required substantial dimensionality reduction, which could obscure or discard such

1 An emerging econometrics literature distinguishes between prediction-policy and causal inference research, maintaining that in the latter context the most appropriate role for computational tools, including large language models, often is to create measurable variables from unstructured data or otherwise generate inputs for downstream statistical analysis (see Kleinberg et al. 2015; Ludwig et al. 2025).

complex attributes. Second, they enable the extraction of relatively complex features that earlier machine learning methods could capture only at the cost of assembling extensive, task-specific training datasets. By reducing or eliminating that requirement, LLMs lower the barriers to incorporating new, previously cost-prohibitive dimensions into empirical legal analysis. Overall, the promise of LLMs in this area appears substantial enough that some have described this application as the "killer app for LLMs in empirical legal research" (Choi 2025).

There are two main approaches to feature extraction with the help of LLMs. One combines LLMs with traditional supervised learning methods. Snippets of text – such as paragraphs from legal opinions or contractual documents – are first converted into vector representations (embeddings). Embeddings are then paired with human-coded datasets to train more conventional machine learning algorithms, which can subsequently classify unlabeled portions of a corpus. Thalken et al. (2023) and Stiglitz & Thalken (2024) illustrate this approach, using transformer models on annotated Supreme Court opinions to classify modes of legal reasoning. Frankenreiter (2025) applies this strategy to identify fee-shifting and forum-selection provisions in corporate charters and bylaws. By preserving richer linguistic information than bag-of-words approaches, embedding-based methods make it possible to detect features that earlier computational methods often could not classify, in particular those that depend on subtle stylistic or semantic cues.

The second approach relies entirely on LLMs and uses their text-generation capabilities to produce classifications directly. For example, Frankenreiter & Talley (2026) identify the presence of 102(b)(7) waivers in corporate charters by providing the text of charters to OpenAI's ChatGPT, together with a prompt instructing the model to determine whether the provision appears. Extending this method, Frankenreiter & Hirst (2026) show that ChatGPT can generate highly detailed codings of complex provisions (advance notice bylaws), in some cases matching or exceeding the performance of human coders. Similarly, Oliver et al. (2024) demonstrate that ChatGPT can accurately classify the presence of suspicious factors in narrative descriptions of traffic-interdiction stops. Unlike the first approach, this method requires no labeled training data: the classification step draws entirely on the linguistic and contextual patterns LLMs learned during model training, substantially reducing the costs of compiling datasets.

However, the second approach also comes with challenges. In the first approach, techniques like cross-validation allow for a robust assess-

ment of an extraction pipeline's accuracy during the training stage. By contrast, when predictions are generated directly by an LLM, assessing reliability is less straightforward. This difficulty is compounded by the "jagged technological frontier" of AI (Dell'Acqua et al. 2023): LLMs can excel at some classification tasks while failing at others, and performance often defies intuitive expectations about task difficulty. Moreover, outputs can be highly prompt-dependent. Against this backdrop, careful validation against human-coded datasets is essential. To avoid overfitting, these validation datasets should be distinct from any datasets used to refine prompts or otherwise tune the classification process. That said, the scale of human labeling required for validation is often much smaller than for training a separate classifier – often hundreds rather than thousands (or more) of labeled examples.

D. Prediction, Classification, Description

In our previous review, we also highlighted that computational methods are sometimes applied not just to generate datasets for downstream analysis, but to produce results more directly. Prediction tasks – such as forecasting case outcomes – are a central example. Large language models can also be deployed in this mode. Their ability to process and generate legal text in contextually rich, human-like ways, and to adapt across diverse legal domains, makes them well-suited to support novel forms of legal prediction, from simulating judicial reasoning to generating plausible arguments for each side in a dispute. An example of this approach is Nigam et al. (2024), who employ LLMs to predict the outcomes of court cases based on the fact patterns presented.

Beyond text generation, researchers have begun to use the vector representations produced by these models to obtain measures for the meaning of words in legally relevant contexts. For example, Nyarko & Sanga (2022) develop a statistical test that leverages word embeddings to quantify differences in meaning across groups and contexts. Even more ambitiously, Arbel & Hoffman (2024) introduce a method for estimating the ordinary meaning of contract terms, and Choi (2024) applies a related embedding-space methodology to quantify uncertainty in statutory language.

While this line of work – particularly the ability to leverage the language-generating capabilities of LLMs to produce novel forms of pre-

dictions – holds great promise, it also faces major challenges. The most important concerns the validation of these tools to ensure they produce meaningful output. This problem is related to the one described in the previous section. LLMs are built on the basis of next-word-prediction technology, and their ability to generate legally meaningful outputs is not a given. In other words, there is an alignment problem: LLMs are optimized for next-word prediction, not for producing legally correct answers, and so their actual performance in this domain requires rigorous validation. Unlike in the context of feature extraction tasks, which usually allow for straightforward accuracy tests with validation data, such validation can be complicated. Especially when LLM use involves the generation of unstructured text or when embeddings are used to quantify the meaning of terms, it can be difficult to assess the extent to which the generated text or measures are meaningful or correct (see Tobia 2024).

E. Engineering and Benchmarking Studies

The emergence of LLMs has spawned a new category of legal AI research that approach legal tasks as engineering challenges to be systematically benchmarked and optimized. This engineering-focused approach prioritizes measurable performance improvements over theoretical insights about law itself. Unlike the causal inference and prediction studies discussed in earlier sections, which seek to understand legal phenomena, engineering studies focus primarily on optimizing model performance against standardized metrics.

One significant development in this line of research has been the creation of large-scale benchmarks designed to evaluate LLMs across multiple dimensions of legal reasoning. LegalBench (Guha et al. 2023), developed through a collaborative process involving over 40 contributors, consists of 162 tasks covering six different types of legal reasoning. LaborBench (Hariri & Ho 2025) is a dataset based on state unemployment insurance laws to assess performance on extracting statutory information. Zheng et al. (2025) develop a benchmark based on bar exam questions to assess performance on retrieval augmented generation tasks in the legal domain. LEXam (Fan et al. 2026) focuses on long-form reasoning derived from 340 actual law exams comprising 4,886 questions in English and German. LexEval represents China's largest legal evaluation dataset, introducing standardized comprehensive benchmarks for evalu-

ating LLMs across Chinese legal contexts (Li et al. 2024). These efforts highlight the importance of multilingual and cross-jurisdictional evaluation frameworks as LLMs are deployed globally.

A parallel strand of engineering research focuses on developing specialized legal LLMs through targeted training and fine-tuning approaches. Research by Niklaus et al. (2024) demonstrates that law-specific instruction training can improve LLM performance on legal task benchmarks. Studies have explored various approaches to legal model optimization, from domain-specific pre-training to instruction fine-tuning on curated legal corpora (Dominguez-Olmedo et al. 2024). Related research has examined prompt engineering strategies (Zambrano 2024) and few-shot learning approaches (Doyle & Tucker 2024). Others have focused on semantic role extraction and legal information extraction tasks (Bakker et al. 2025; de Faria et al. 2025).

One pressing engineering challenge for the practical use of LLMs in the legal context has been hallucinations – instances where LLMs generate plausible but factually incorrect legal information. Dahl et al. (2024) revealed that hallucinations occur between 58 % of the time with ChatGPT-4 and 88 % with Llama 2 when asked specific, verifiable questions about federal court cases. Subsequent research by Magesh et al. (2025) evaluated commercial legal research tools, finding that Lexis+ AI and Westlaw AI-Assist hallucinate between 17%-33 % of the time despite using retrieval-augmented generation techniques. This work demonstrates that even sophisticated engineering approaches cannot entirely eliminate hallucinations, emphasizing the need for transparent benchmarking of commercial products.

F. Challenges and Limitations

While large language models offer transformative capabilities for legal analysis, their integration into legal scholarship and practice poses a number of important challenges that must be carefully addressed. Among the most pressing concerns are accuracy and reliability. Unlike traditional computational methods where errors are typically systematic and predictable, LLM failures can be subtle, inconsistent, and difficult to detect. Plausible hallucinations and the jagged technological frontier phenomenon can make it difficult for legal scholars and practitioners to understand and predict errors associated with LLMs. Legal change may also

prove to be a fundamental limit on the use of LLMs in certain cases. Legal information evolves continuously through new legislation, regulations, and case law. LLMs trained on historical data are much more well-suited for providing a snapshot based on prior information than offering clear guidance on the current state of the law.

Another important set of challenges concern bias and fairness. Because LLMs are trained on vast corpora of legal texts that reflect historical patterns of legal decision-making, their outputs may encode prior biases or discriminatory behavior. This fact can help reveal patterns in ways that are useful for researchers (Ash and Chen), but scholars must work carefully to ensure that their work illuminates, rather than perpetuates, bias. The scale and opacity of LLM training data make bias identification and mitigation particularly challenging. Unlike traditional machine learning approaches where training datasets can be systematically audited, LLM training involves processing billions of documents whose contents and biases cannot be comprehensively characterized.

The integration of LLMs into legal practice also raises fundamental questions about professional responsibility that existing ethical frameworks are not equipped to address. Legal professionals have duties of competence, confidentiality, and zealous advocacy that may be compromised by inappropriate reliance on LLM systems. Courts and regulatory bodies are grappling with how to address LLM-generated content in legal proceedings, raising novel questions about admissibility, authentication, and reliability standards. Traditional rules of ethics, evidence, and procedure were not designed to handle content generated by artificial intelligence systems, creating uncertainty.

G. Conclusion

The emergence of large language models represents a paradigm shift in how legal scholars conduct empirical research. Our review identifies three dominant applications: data generation and feature extraction that dramatically reduce barriers to incorporating complex textual features; inference and prediction applications unlocking new forms of legal analysis; and engineering studies establishing technical foundations despite persistent accuracy challenges. These developments collectively represent both the culmination of decades of progress and the opening of entirely new research frontiers.

However, this transformation brings fundamental research challenges that demand sustained scholarly attention. The persistent problem of hallucinations introduces validation complexities absent in traditional computational methods. Models may excel at sophisticated legal reasoning while failing unpredictably at seemingly simpler tasks, defying intuitive assessment of reliability. Beyond technical limitations, LLMs raise deeper methodological questions about the nature of legal knowledge and reasoning. When models trained on vast legal corpora produce insights that match or exceed human expert performance, scholars must grapple with the difference between genuine understanding or sophisticated pattern matching, and what this distinction means for empirical legal research.

Moving forward, legal scholars should approach LLMs as powerful tools requiring careful validation rather than as replacements for traditional methods. In the domain of practice, professions should view these tools as sophisticated assistants that require close oversight, particularly given hallucination risks. And policymakers face the challenge of balancing LLMs substantial benefits with requirements for accuracy, fairness, and accountability through new validation standards and transparency guidelines. The response of these communities will determine how these recent technological innovations will affect the practice and study of the law.

References

Arbel, Yonathan, and David A. Hoffman. "Generative Interpretation." *New York University Law Review* 99, no. 2 (May 2024): 451–514.

[Ash, same volume].

Bakker, Roos, Akke Schroevers, Romy van Drie, et al. "Semantic Role Extraction in Law Texts: a Comparative Analysis of Language Models for Legal Information Extraction." *Artificial Intelligence and Law* (2025).

Choi, Jonathan. "Measuring Clarity in Legal Text." *University of Chicago Law Review* 91, no. 1 (2024): 1–82.

Choi, Jonathan. *A Highly Tactical Guide to Generating Data from Legal Text with LLMs*. Unpublished manuscript, 2025.

Dahl, Matthew, Varun Magesh, Mirac Suzgun, and Daniel Ho. "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models." *Trademark Reporter* 114, no. 6 (2024): 880–909.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, et al. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on*

Knowledge Worker Productivity and Quality. Harvard Business School Working Paper No. 24–013, September 2023.

Dominguez-Olmedo, Ricardo, Vedant Nanda, Rediet Abebe, et al. "Lawma: The Power of Specialization for Legal Annotation." In *Proceedings of the Thirteenth International Conference on Learning Representations* (ICLR 2025), 35059–35103. Appleton, WI: ICLR, 2025.

Doyle, Colin, and Aaron Tucker. "If You Give an LLM a Legal Practice Guide." In *Proceedings of the 2025 Symposium on Computer Science and Law* (CSLAW '25), 194–205. New York: Association for Computing Machinery, 2025.

Fan, Yu, Jingwei Ni, Jakob Merane, et al. "LEXam: Benchmarking Legal Reasoning on 340 Law Exams." *arXiv preprint* arXiv:2505.12864 (January 2026).

De Faria, Joana Ribeiro, Huiyuan Xie, and Felix Steffek. "Information Extraction from Employment Tribunal Judgments Using a Large Language Model." *Artificial Intelligence and Law* (2025). https://doi.org/10.1007/s10506-025-09443-z.

Frankenreiter, Jens. "The Other Delaware Effect." *Washington University in St. Louis School of Law Legal Studies Research Paper* No. 25-03-11, 2025. https://ssrn.com/abstract=5115285.

Frankenreiter, Jens, and Scott Hirst. *The Rise of the Horse-Choker: The Complexification of Advance Notice Bylaws*. Unpublished manuscript, 2026.

Frankenreiter, Jens, and Eric Talley. "Sticky Charters? The Surprisingly Tepid Embrace of Officer-Protecting Waivers in Delaware." *Harvard Business Law Review* 16 (2026): 105–168.

Frankenreiter, Jens, and Michael Livermore. "Computational Methods in Legal Analysis." *Annual Review of Law and Social Science* 16 (2020): 39–57.

Guha, Neel, Julian Nyarko, Daniel Ho, et al. "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (NeurIPS '23), 44123–44279. Red Hook, NY: Curran Associates, Inc., 2023.

Hariri, Emaan, and Daniel Ho. "AI for Statutory Simplification: A Comprehensive State Legal Corpus and Labor Benchmark." In *Proceedings of Twentieth International Conference on Artificial Intelligence and Law* (ICAIL '25), 177–187. ACM, New York. 2025.

Janatian, Samyar, Hannes Westermann, Jinzhe Tan, et al. "From Text to Structure: Using Large Language Models to Support the Development of Legal Expert Systems." *arXiv preprint* arXiv:2311.04911 (November 2023).

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. "Prediction Policy Problems." *American Economic Review* 105, no. 5 (2015):491–95.

Li, Haitao, You Chen, et al. "LexEval: A Comprehensive Chinese Legal Benchmark for Evaluating Large Language Models." In *Proceedings of the Thirty-Eighth Conference on Neural Information Processing Systems* (NeurIPS 2024), 25061–25094. Red Hook, NY: Curran Associates, Inc., 2024.

Livermore, Michael, and Daniel Rockmore, eds. *Law as Data: Computation, Text, & the Future of Legal Analysis*. Santa Fe Institute Press, 2019.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. "Large Language Models: An Applied Econometric Framework." *arXiv preprint* arXiv:2412.07031 (December 2025).

Magesh, Varun, Faiz Surani, Matthew Dahl, et al. "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools." *Journal of Empirical Legal Studies* 22 (2025): 216–242.

Nigam, Shubham Kumar, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. "Rethinking Legal Judgement Prediction in a Realistic Scenario in the Era of Large Language Models." In *Proceedings of the Natural Legal Language Processing Workshop 2024*, 61–80. Miami, FL: Association for Computational Linguistics, 2024.

Niklaus, Joel, Elliott Ash, Matthias Stürmer, and Ilias Chalkidis. "LawInstruct: A Resource for Studying Language Model Adaptation to the Legal Domain." *arXiv preprint* arXiv:2404.02127 (April 2024).

Nyarko, Julian, and Sarath Sanga. "A Statistical Test for Legal Interpretation: Theory and Applications." *Journal of Law, Economics, & Organization* 38, no. 2 (2022): 539–569.

Oliver, Wesley, Morgan Gray, Jaromir Savelka, and Kevin Ashley. "Computationally Assessing Suspicion." *University of Cincinnati Law Review* 92, no. 4 (2024): 1108–1170.

Stiglitz, Edward, and Rosamond Thalken. "Historical Trends in Macro-Jurisprudence: A Language Model Assessment, 1870–2023." *Maryland Law Review* 84, no. 1 (2024): 46–101.

Thalken, Rosamond, Edward Stiglitz, David Mimno, and Matthew Wilkens. "Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9252–9265. Association for Computational Linguistics, December 2023.

Tobia, Kevin. "Algorithmic Interpretation: A Response to Professor Jonathan Choi's *Measuring Clarity in Legal Text*." *University of Chicago Law Review Online*, 2024.

Zambrano, Guillaume. "Case Law as Data: Prompt Engineering Strategies for Case Outcome Extraction with Large Language Models in a Zero-Shot Setting." *Law, Technology and Humans* 6, no. 3 (2024): 80–101.

Zheng, Lucia, Neel Guha, Javokhir Arifov, et al. "A Reasoning-Focused Legal Retrieval Benchmark." In *Proceedings of 4th ACM Symposium on Computer Science and Law* (CSLAW '25). New York: Association for Computing Machinery, 2025.