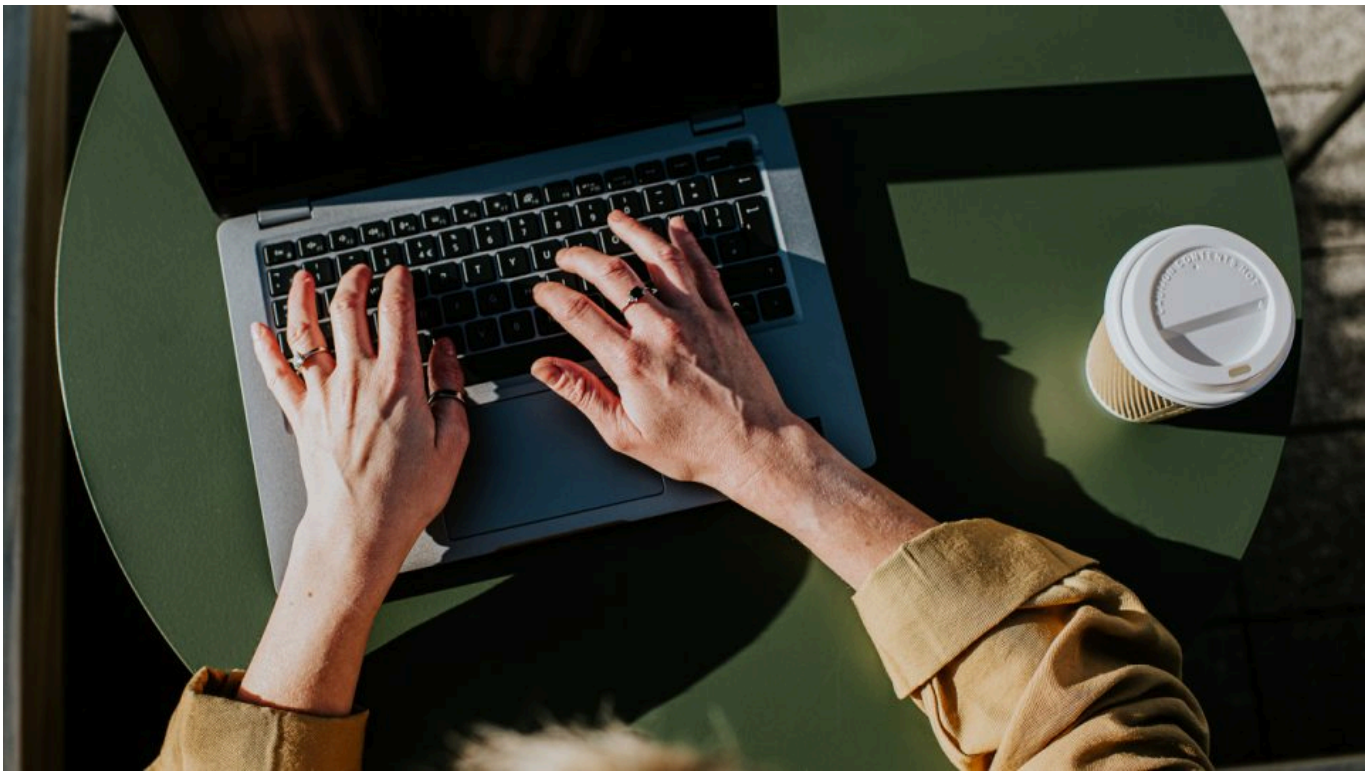
How Large Language Models (LLMs) Are Transforming Litigation Workflows

Michael Aiken and Derek Riley

Oct 14, 2025 10 min read

Summary

- Large language models automate legal research, discovery, and document drafting, increasing efficiency and accuracy in litigation workflows.
- Retrieval-Augmented Generation (RAG) integrates legal-specific data, reducing errors and enhancing the reliability of AI-generated legal outputs.
- Ethical use of AI in law requires human oversight, validation of outputs, and adherence to professional responsibility standards.



Catherine Falls Commercial via Getty Images

Introduction

Large language models (LLMs) such as OpenAI's GPT-4, Google Gemini, Meta's LLaMA, and Anthropic's Claude are revolutionizing the legal industry. Built on transformer-based deep learning architectures, these artificial intelligence (AI) systems can analyze, summarize, generate, and reason across vast bodies of text – capabilities that align closely with the language driven nature of legal work. Originally trained on general-purpose internet-scale collection of digital text (*i.e.* corpora), foundational and specialized LLMs are now able to serve specific industries, including law.

For the legal profession, LLMs offer immense promise, as both operate in a world of language, context and meaning. Rather than replacing attorneys, LLMs serve as powerful tools to augment legal expertise, automate repetitive tasks, and streamline complex workflows. Law firms that embrace this technology will gain a significant competitive advantage, particularly in high-volume or high-analysis areas like discovery, legal research, briefing, and trial preparation. As clients increasingly expect efficiency and value, leveraging LLMs will become as essential as using modern word processors or legal databases.

This whitepaper explains technically how LLMs work, how they can be augmented with legal-specific data via Retrieval-Augmented Generation (RAG), the fundamentals of prompt engineering, the ethical and professional obligations surrounding their use, and the marketplace of tools currently shaping the legal AI ecosystem.

Understanding LLMs: Technical Foundation for Legal Professionals

Legal professionals need to understand the technical foundation for LLMs: the transformer architecture that empowers nuanced legal reasoning and drafting; the vast, diverse datasets that fuel LLMs' capabilities—while also exposing their blind spots and risk of "hallucinations" without domain-specific context; the "black box" nature of AI decision-making, underscoring the necessity of human oversight; and the practical realities of tokenization and context windows, which shape how LLMs process lengthy legal documents and demand strategic input management. By mastering these fundamentals, legal practitioners will be equipped to harness LLMs' power effectively and responsibly in their workflows.

1. Generative Pre-trained Transformers

LLMs such as GPT-4 are built upon the transformer architecture introduced by Vaswani et al. (2017). [1] Although the transformer architecture is a concept going back decades, successful training of LLMs were enabled by recent developments in computing power (think NVIDIA) to train on extremely large data sets. These models rely on a mechanism called self-attention, which enables the LLM to weigh the relationships between various elements of the input text within a context window, allowing for nuanced responses considering context, syntax, and semantics. This allows the model to perform a wide range of language-based tasks, from summarization and translation to legal analysis and argumentation. Ultimately, LLMs are next-token prediction models, so they generate essentially one word at a time and are able to produce large amounts of syntactically correct responses with semantic validity.

2. Training on Massive Datasets

LLMs are pre-trained on diverse text corpora including legal cases, government filings, websites, books, and other digital documents. While these corpora are vast, they are not exhaustive, and access to proprietary or up-to-date legal databases is not included in publicly available foundation models that are used by popular LLM services. Consequently, general-purpose LLMs will not possess relevant domain knowledge for many tasks, which can lead to responses that contain semantic errors, often called hallucinations. This is akin to a law student who has studied every proceeding, brief, and manuscript in a law library but is asked to write a summary of a document they haven't seen.

To overcome the problem of hallucinations critical domain information can be added to the prompt in the context window to give the model a concrete foundation for its responses. Systematic approaches to do this across specialized document databases exist such as retrieval-augmented generation; however, these approaches require nuance and tuning for the domain to operate accurately. Additionally, context window sizes are limited to hundreds of pages of text, so for domains where more specialization is necessary or style adaptation is required, fine-tuning foundation LLMs is an option.

3. The Black Box Problem

While outputs from LLMs can be highly syntactically and/or semantically accurate, the internal mechanisms of the LLM that were used to lead to the output remain largely opaque—a phenomenon referred to as the "black box" problem. [2] The model's decisions result from complex internal representations that are difficult to trace, and while LLMs can be asked to justify their output, there are no guarantees that the

justifications are accurate or representative of the underlying decisions that are made internally in the LLM.

Empirical evaluations have shown LLMs can outperform junior attorneys on multiple-choice legal reasoning tasks and contract drafting assessments. [3] This is possible due to the sheer quantity of data these models have contextualized as well as the capacity of the models, measured in their size, which can be on the order of Trillions of learnable parameters. Confidence in the quality of output can be easily built through using these systems, but care must always be taken with the output to validate the accuracy.

4. Tokens and Context Windows

Rather than processing words, LLMs break inputs into tokens, which are typically short character sequences. Each model has a token limit, or "context window," which determines the amount of text it can process in a single session. GPT-4 Turbo, for example, supports up to 128,000 tokens, which is approximately the length of a 300 page novel. The latest LLMs can have context windows that support more than a million tokens, but LLMs may not pay attention to everything in the context window. Research has shown that LLMs tend to ignore context in the middle of the context window, preferring to pay attention to context toward the beginning and end of the context window.

Context windows have significant implications for uploading case files, deposition transcripts, or statutes in full. If the input is too long it may be cut off at the beginning or end, leading to missing key information or inaccurate responses. In addition, with full legal filings or a long transcript, the model might focus on the most recent or prominent parts or miss context spread throughout the document. Legal professionals should be aware of context window limits and use techniques like chunking, embeddings,and summarization to manage long inputs.

Integrating Legal-Specific Data: Retrieval-Augmented Generation

General-purpose LLMs excel at reasoning but often lack specialized knowledge required for legal practice. Fine tuning LLMs on proprietary data is possible but resource-intensive and redundant. Instead, most legal AI tools now use Retrieval-Augmented Generation (RAG) to inject domain-specific content into the model context.

1. What Is RAG?

RAG combines an LLM with a high-performance database of relevant legal documents. When a user submits a query, the system retrieves pertinent documents (e.g., contracts, filings, depositions) and feeds them into the LLM's context window. This approach grounds the model's responses in authoritative sources without retraining the model, enhancing bot accuracy and security.

2. Chunking and Vectorization

Before documents can be used in RAG, they are divided ("chunked") into logical segments and converted into vector embeddings for efficient semantic search. Only the most relevant chunks are provided to LLM reducing irrelevant details and improving the quality of the responses. The chunking strategy, embedding model, and retrieval algorithm all critically influence RAG performance and the legal usefulness of responses. [4] This is a differentiating feature of emerging software applications serving the legal field and are critical to both limiting hallucinations (fabricated information) and providing verifiable citations.

3. Model Plug-and-Play

Because RAG separates the knowledge base from the language model, newer or better LLMs can be swapped into existing pipelines without re-architecting the system. This modularity ensures legal organizations stay up to date with the latest models while maintaining their proprietary knowledge infrastructure. Eventually, companies with specialized LLM applications will submit benchmarking of their use of various foundational models to guarantee that they are getting optimum performance for the task at hand, to differentiate their product. Indeed, the foundational models have already performed benchmarking against each other on legal reasoning. [5]

Incorporating LLMs into Legal Workflows

Lawyers using LLM technology must adhere to both ethical obligations and practical concerns. [6] Attorneys are bound by the ABA Model Rules, including Rule 1.1 (Competence), which now includes technological competence, and Rule 1.4(a)(2) regarding client communication. Lawyers must understand "the benefits and risks associated with relevant technology" to comply with these rules, and communicate their intentions with regard to their use of such technology in the scope of the representation.

While LLMs generate high-quality drafts and summaries, they may occasionally produce hallucinations or misstate legal principles. Just like obtaining information from a google

search, lawyers must never rely blindly on AI outputs and must always validate outputs through manual review. Courts have sanctioned attorneys numerous times for submitting filings based on AI-generated content without verification. [7] This human-in-the-loop approach ensures ethical compliance, minimizes risk, and maintains accountability. [8] Further, the software applications should make checking citations much more efficient than a traditional file.

Common Use Cases

LLMs are reshaping litigation workflows by streamlining and enhancing a wide range of legal tasks that traditionally require substantial time and manual effort. They can rapidly draft legal memos, research briefs, and demand letters, reducing the burden on attorneys and increasing consistency across documents. LLMs are also adept at summarizing complex depositions and organizing key evidence, enabling faster and more effective case analysis. In discovery, they assist in preparing detailed and tailored requests or responses. LLMs can generate structured outlines for direct and cross-examinations, helping attorneys stay focused on critical themes. They offer advanced tools like sentiment analysis to aid in evaluating witness credibility, and can identify inconsistencies in testimony across multiple depositions. Additionally, LLMs facilitate contract review and redlining by quickly flagging problematic clauses or deviations from standard terms. Perhaps most strategically, they support legal argument development by helping attorneys craft persuasive narratives, analogies, and themes, contributing to more compelling briefs and advocacy.

Prompt Engineering: Getting the Best from LLMs

The quality of LLM output depends heavily on the quality of the prompt. [9] Prompt engineering refers to the structured development of instructions given to an LLM to guide it toward a desired outcome. Generally, you should be maximally prescriptive, as word choice, style, tone, structure and context of the prompt all matter.

Effective prompts are:

- **Specific.** Clearly state the task.
- **Context-Rich.** Provide background, tone, and format.
- **Structured.** Explicitly define the expected output.

For example, this is an effective prompt:

> Task: Identify all the basis for impeachment from the deposition of Mr. Smith.
>
> Context: Mr. Smith is a defendant in breach of fiduciary duty case.
>
> Output: Create a list of all potential areas of impeachment, with citations to supporting testimony from other witnesses and/or documents. Constrain your output to what is supported in the transcripts and documents provided.

In addition, it is often recommended to assign the LLM an explicit role. For example, you are a highly analytical trial attorney preparing for cross-examination of Mr. Smith.

Researchers recommend zero-shot (no examples), few-shot (2–6 examples), and chain-of-thought prompting (ask for step-by-step reasoning) depending on task complexity. [10] Techniques such as step-back prompting (zooming out to consider context) and prompt self-evaluation (asking the model what it found unclear) are also emerging best practices.

In addition, there are techniques like chain of thought prompting to ask the LLM to explain its reasoning step by step or going through the steps with individual questions; "step back" prompting where you aske the LLM to consider a more general knowledge question first to activate relevant background information, and then consider a specific application; and automated prompting to generate prompts by using the LLM, asking how the prompt could be improved, or if there is anything that is unclear. Prompting is a naturally iterative process. Some software includes prompt histories and prompt libraries for common tasks.

Finally, models also include sampling controls like temperature to impact the extent of randomness in the output versus being deterministic in terms of the weightings. More quantitative prompts should have a low temperature, while more creative output could be set higher temperature. Generally, you will want the model in a low temperature mode for legal work.

LLMs in the Legal Software Marketplace

A rapidly growing marketplace of legal AI tools is leveraging foundational LLMs to offer specialized capabilities:

- **Alexi.com** – AI powered legal research, document summarization, and litigation prep.
- **Briefpoint.ai** – Automated discovery drafting for interrogatories and RFAs.
- **Callidus.ai –** AI powered legal research and drafting.

- **Clio.com** – AI-integrated practice management software.
- **Iqidis.ai** – Personalized AI tailed to lawyer's identity and practice with RAG system to inspect and edit citation sources. Promises at least 50% time saved on legal tasks. Performs legal research, drafting and analysis.
- **Filevine.com** – AI-enhanced case management, document information extraction, organization, deposition analysis, and calendaring. Performs data mapping and extracts key information. Can generate demand letters in minutes with supporting facts. Custom prompt builder.
    - **Skribe.ai / Depo Copilot** – Deposition recording, summarization and analysis.
- **Harvey.ai** – End-to-end legal assistant built on OpenAI, tailored for elite law firms. High volume document analysis to extract key information and summarize. Specializes in legal brief writing.
- **Lawme.ai** – Suite of AI powered tools on client onboarding, bulk data extraction, legal research and contract drafting.
- **LawLM.ai** – Focus on AI summaries and analysis of transcripts with citations to page and line, and RAG chatbot to analyze testimonial evidence across multiple witnesses on large complex cases to aid with preparation for discovery and trial.
- **Lexis+** –proprietary LLM and tool suiteAlso features AI assisted case law research with Lexis content.
- **TryNovo.com** – Tools for demand letters and medical chronologies.
- **Paxton.ai –** AI-powered legal research and drafting platform, designed to help attorneys and law firms streamline tedious tasks and enhance productivity. Confidence indicator and AI citator.
- **SmartAdvocate.com** – case management with AI tools to summarize cases, briefs and records, and other integrated non-AI features.
- **Supio.com –** specialized AI for plaintiff's personal injury document formatting and data.
- **Spellbook.legal** – GPT-4 contract drafting and review tool.

- **Upcounsel** – Thomson Reuters LLM tool suite that includes analysis of large sets of legal documents and extract key documents and information. Also features AI assisted case law research with Westlaw content.

These tools span legal research, discovery, case management, evidence review, trial preparation, and client engagement. They promise to democratize legal access by enhancing small firm capabilities, allowing lawyers to focus on high-value strategic work, and reducing costs for clients. Customer data is not used to train the models utilized by the applications, and typically data is encrypted in transit and in rest.

Not all tools are created equal. Some are little more than "ChatGPT wrappers" with minimal added value. Look for features like custom databases, RAG integration, citation management, and workflow embedding to distinguish robust solutions from superficial ones.

The Future of LLMs in Legal Practice

Generative AI is still in the early stages of disrupting the legal industry. As LLMs become more capable, secure, and explainable, they will reduce friction in legal workflows, lower barriers to justice, and enable innovative legal strategies. While some large firms may build proprietary models, the prevailing trend is to leverage general-purpose LLMs with custom RAG pipelines.

Ethical considerations, evolving billing models (e.g., reduced reliance on the billable hour), and bar association guidance will shape this adoption. Early adopters will enjoy compounding advantages: better client service, faster onboarding, easier collaboration and greater transparency. Firms that resist this technological shift may find themselves at a competitive disadvantage.

Endnotes

1. Ashish Vaswani et al., *Attention Is All You Need*, in 30 Advances in Neural Info. Processing Sys. (2017).

2. Zachary C. Lipton, *The Mythos of Model Interpretability*, *arXiv* (June 2016), https://arxiv.org/abs/1606.03490.

3. Michael J. Bommarito II & Daniel Martin Katz, *GPT Takes the Bar Exam*, *SSRN Electronic Journal* (2023), https://doi.org/10.2139/ssrn.4314839.

4. Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, in 33 Advances in Neural Info. Processing Sys. (2020).

5. Yu Fan et al., *LEXam: Benchmarking Legal Reasoning on 340 Law Exams*, *arXiv* (May 2025), https://arxiv.org/abs/2505.12864; also available at SSRN, https://ssrn.com/abstract=5265144.

6. See ABA Comm. on Ethics & Prof'l Responsibility, Formal Op. 512 (2023), https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf.

7. See, e.g., *Lawyers in Walmart Lawsuit Admit AI Hallucinated Case Citations*, Reuters (Feb. 10, 2025), https://www.reuters.com/legal/legalindustry/lawyers-walmart-lawsuit-admit-ai-hallucinated-case-citations-2025-02-10/.

8. L. Boonstra, H. Sherman & Y. Cao, *Prompt Engineering* (Google White Paper, Sept. 2024), https://storage.googleapis.com/cloud-ai-blog-public/docs/ptompt-engineering-whitepaper.pdf.

9. Jason Wei et al., *Chain of Thought Prompting Elicits Reasoning in Large Language Models*, *arXiv* (Jan. 2022), https://arxiv.org/abs/2201.11903.

10. Nelson F. Liu et al., *Lost in the Middle: How Language Models Use Long Contexts*, *arXiv* (July 2023), https://arxiv.org/abs/2307.03172.