The State of Large Language Models in August 2025

Seth J. Chandler, with help from AI

August 15, 2025

1 Introduction: Beyond the Chatbot

Welcome. If you are reading this, you have enrolled in this course because you sense, correctly, that the technology of Large Language Models (LLMs) represents something more than just a clever chatbot for drafting emails or summarizing articles. You have likely experimented with ChatGPT, Gemini, or Claude, and have been alternately impressed by their fluency and frustrated by their limitations. The prevailing public perception of this technology is frozen in its 2023-2024 iteration: a useful but often unreliable tool, prone to making things up, and fundamentally incapable of genuine "thinking." The purpose of this report, and indeed this course, is to disabuse you of that simplistic notion. It is to stop you from thinking that the basic, consumer-facing use of these tools is all there is.

That view is a canard. It is akin to looking at an eight-month-old child, observing that it can neither walk nor talk, and concluding it will never amount to anything. The field of generative AI is not moving at a linear

pace; its capabilities are compounding at a rate that is difficult for even seasoned practitioners to fully internalize. The universe of LLMs in August 2025 is a vast, expanding, and deeply complex ecosystem. It is a world of multi-trillion parameter models, of open-source communities rivaling corporate giants, of novel interaction paradigms that transcend simple prompting, and of strategic imperatives that are already reshaping the business of law.

This report will serve as your initial survey of this new landscape. We will explore the current capabilities and limitations of the frontier models, moving past the myths that cling to their earlier, less capable ancestors. We will discuss the innovations that define the current state-of-the-art: the ubiquity of computationally intensive "thinking modes," the evolution of "prompt engineering" into a more intuitive and collaborative practice, and the explosion of open-source models that offer unprecedented power and control. We will also address the persistent and critical issue of "hallucinations"—the generation of plausible but false information—not as an insurmountable flaw, but as a manageable risk that requires new skills and disciplined workflows.

For you, as aspiring legal professionals, understanding this ecosystem is not an elective. It is a core competency. The ability to strategically leverage these tools—and, just as importantly, to understand their failure modes—will soon be as fundamental to the practice of law as the ability to conduct traditional legal research. Your objective in reading this is not merely to learn a new piece of software, but to begin building a new mental model for knowledge work itself. The universe is far larger than you imagine. Let us begin the exploration.

2 The New Frontier: Models and Capabilities in August 2025

The competitive landscape of late 2025 is no longer a simple two-horse race between OpenAI and Google. It is a dynamic, multipolar world, characterized by intense competition at the proprietary frontier, a booming open-source movement providing powerful and customizable alternatives, and a sophisticated layer of intermediary services that democratize access to all of them.

2.1 The Proprietary Titans: GPT-5, Gemini 2.5, and Claude 4

The largest, most capable models remain the products of a few well-funded corporate labs. However, their defining features are no longer just about scale, but about specialized modes of operation and deep integration into workflows.

2.1.1 OpenAI's ChatGPT-5 and the Ubiquity of "Thinking Models"

The release of GPT-5 solidified a trend that had been emerging for over a year: the bifurcation of model interaction into distinct modes. The default, consumer-facing mode of ChatGPT-5 is a "flash" or "chat" model, optimized for speed, efficiency, and low-cost inference. It is remarkably capable for everyday tasks: summarizing text, drafting correspondence, and answering straightforward factual questions.

The true innovation, however, is its "thinking" or "Opus" mode. This is not a different model, but rather a different way of running the same model. When a user activates this mode, the system allocates significantly more computational resources—or "compute"—to the query. Instead of generating a response in a single, rapid pass, the model engages in a form of internal monologue or multi-step reasoning. It might generate and critique several potential answers, perform a chain-of-thought process that is orders of magnitude more complex than before, and verify its own reasoning against different internal frameworks before producing a final output.

The result is a qualitative leap in analytical depth. For a lawyer, this is the difference between asking an associate for a quick off-the-cuff answer versus asking for a thoroughly researched memorandum. The "thinking" mode can analyze a complex fact pattern against a novel legal theory, identify subtle inconsistencies in an opponent's brief, or draft a contractual clause with multiple, nested dependencies, all with a degree of coherence that was previously unattainable. The trade-off is, of course, time and cost. A query in "thinking" mode might take thirty seconds to a minute to process and cost ten times as much as a standard query. But for high-stakes legal work, this trade-off is not just acceptable; it is essential. This dual-mode architecture is now standard across all frontier models, including Gemini and Claude.

2.1.2 Google's Gemini 2.5 and the "Study and Learn" Paradigm

Google's Gemini series has continued to leverage its unique advantage: deep integration with a vast ecosystem of user data and services. While privacy concerns remain a constant point of negotiation, the utility is undeniable. The most significant innovation in Gemini 2.5 is the maturation of its "Study and Learn" mode.

This mode transforms the LLM from a simple question-answering machine into an interactive, Socratic tutor. A law student can, for example, upload their entire syllabus for Civil Procedure, along with class notes and assigned readings. They can then ask Gemini not just to summarize the concept of *res judicata*, but to act as a study partner. The student can ask the model to generate practice hypotheticals, to critique their written analysis of a case, or to engage them in a dialogue about the policy implications of the Federal Rules.

Crucially, the "Study and Learn" mode maintains a persistent state of the user's progress. It knows which concepts the student has struggled with and can tailor future interactions to reinforce those areas. It can adapt its level of explanation from a simple overview to a graduate-level discourse. This represents a profound shift from a transactional model of interaction (one prompt, one answer) to a relational one, where the AI becomes a personalized tool for cognitive enhancement and skill acquisition over time. Claude has a similar, though less deeply integrated, version of this feature, and it is rapidly becoming a key differentiator in the educational and professional development markets.

2.2 Other Notable Models in the Ecosystem

Beyond the "big three" of OpenAI, Google, and Anthropic, the LLM landscape is populated by other models that, for a lawyer, are more notable for their limitations than their strengths. While they may serve niche purposes or provide a no-cost alternative, they do not currently compete with the frontier models for serious legal work and should be approached with significant caution.

2.2.1 Meta.ai: The Free but Flawed Alternative

Meta's AI, powered by its Llama models, is perhaps the most widely accessible AI on the planet, integrated directly into Facebook, Instagram, and WhatsApp. For any serious legal application, however, it is a profoundly flawed tool. Its responses to legal queries are frequently shallow, riddled with hallucinations, and reflect a level of capability that feels dated—a relic of 2023 in a 2025 world. There is a reason Meta is aggressively trying to hire top AI talent: the current product is simply not competitive for professional use cases. Its only two redeeming qualities are that it is free and fast. You get bad answers quickly, which is rarely a desirable feature in the legal profession. It should not be considered a primary tool for any substantive work.

2.2.2 Grok-4: A Runner-Up with Niche Strengths

Grok-4 from xAI is a significant improvement over its crummy predecessor, but it remains a clear runner-up in the current market. For legal queries, the quality of its answers is still a step behind what GPT-5 and Gemini consistently provide. Furthermore, it lacks the rich ecosystem of features that make the leading models so versatile, such as the ability to create or use customized helpers like CustomGPTs or Gems. While Grok is less "censored" than its competitors and possesses respectable capabilities in mathematics and coding, these strengths do not compensate for its core deficiencies in legal reasoning and its lack of a mature feature set. For a lawyer, it is a secondary or tertiary option at best, not a foundational component of a professional AI toolkit.

2.3 The Chinese Open-Source Revolution

Perhaps the most significant development of the past year has not come from Silicon Valley, but from a distributed, global community of researchers and developers, with major contributions from Chinese AI labs like 01.AI, Zhipu AI, and DeepSeek. Models like the DeepSeek series and Llama 3.1 (from Meta) now exist in sizes ranging from 20 billion to over 500 billion parameters, and their performance on key benchmarks is beginning to rival the proprietary models of just 12-18 months prior.

The importance of this cannot be overstated. For a law firm, the existence of a high-quality, open-source model means they are no longer beholden to the terms of service, privacy policies, or pricing structures of a handful of large tech companies. A firm can download a model like 'DeepSeek-500B', install it on its own private servers (either on-premise or in a secure cloud), and have a powerful AI that has never seen, and will never see, the public internet or any other client's data. This provides an unparalleled level of security and confidentiality for sensitive client information. Furthermore, as we will discuss, it opens the door to fine-tuning—the process of further training a model on a firm's own private data to create a truly bespoke, expert legal assistant.

There is a political and security concern, however, with use of the Chinese models. Some people have great concerns about the role of the Chinese Communist Party in models hosted in or developed in China. They fear that data may not be secure and could be used for blackmail, plagiarism or various forms of hacking. Indeed, the State of Texas bars Texas state employees from using DeepSeek on university-owned hardware. Whether the security threats posed by Chinese models is significantly greater than those posed by use of American models is a matter of both concern and debate.

3 The Evolution of Human-AI Interaction

The way we "talk" to these models is changing just as quickly as the models themselves. The era of arcane, highly technical "prompt engineering" is giving way to more fluid, intuitive, and powerful methods of collaboration.

3.1 From Prompt Engineering to "Vibe Coding"

In the early days of LLMs, getting a good output was a dark art. Users, often called "prompt engineers," would craft long, convoluted prompts filled with specific instructions, role-playing scenarios, and formatting commands. It was a rigid, command-based interaction.

The frontier models of 2025 are far more adept at understanding intent. The practice has evolved into something more akin to art direction or what some have termed "Vibe Coding" or "Vibe Lawyering." Instead of giving a precise command, the user provides a high-level description of the desired *vibe* or *intent*. A lawyer might provide a draft clause and say, "This is too aggressive. Redraft it with a more collaborative tone, but ensure our client is still fully indemnified against claims of gross negligence. The vibe should be 'firm but fair partner.' "

The AI, particularly in its "thinking mode," can now parse this complex, nuanced instruction. It understands the legal concept of indemnification, the stylistic difference between "aggressive" and "collaborative," and the abstract persona of a "firm but fair partner." It will then generate a revised clause that reflects this high-level guidance. This is a profoundly different type of interaction. It is not engineering; it is a conversation. It allows the legal professional to focus on high-level strategy and judgment, using

the AI as a tireless, fluent associate to handle the mechanical aspects of drafting and revision. This paradigm is particularly powerful in Claude's code generation capabilities and is rapidly becoming the standard for all sophisticated drafting tasks.

3.2 API Access and the Rise of Intermediaries

Simultaneously, accessing these powerful models has become dramatically easier for developers and institutions. Every major model is available via an Application Programming Interface (API), allowing it to be integrated into other software. A law firm could, for instance, build a tool directly into their document management system that allows any lawyer to select a document and ask a powerful model to summarize it, compare it to another document, or check it for specific clauses.

This ease of access has been supercharged by the development of intermediary services like OpenRouter. An intermediary acts as a universal translator and switchboard for hundreds of different AI models. Instead of a developer needing to write separate code to connect to GPT-5, then to Claude, then to a dozen different open-source models, they can write a single piece of code that connects to the intermediary. The intermediary then routes the request to the best—or most cost-effective—model for the job.

This has two profound consequences. First, it commoditizes the base models. A firm is no longer locked into a single provider. They can dynamically switch between models based on price, performance, or the specific task at hand. Second, it fuels innovation by allowing small developers to build powerful applications that leverage the best of the entire AI ecosystem without prohibitive complexity.

4 The Persistent Specter of Hallucination

No discussion of LLMs is complete without addressing their most notorious failure mode: hallucination. A hallucination is the generation of information that is plausible, fluent, and contextually appropriate, but factually incorrect or entirely fabricated. Early models were infamous for this, confidently inventing case citations, historical events, and even the number of 'r's in "strawberry." While this problem has been significantly mitigated in the frontier models of 2025, it has not been eliminated. Understanding when and why hallucinations occur is a critical skill for any lawyer using these tools.

4.1 Why Hallucinations Happen and Why They're Getting Better

At their core, LLMs are probabilistic models. They are designed to generate the most likely sequence of words, not to state verified facts. Hallucinations occur when the most statistically probable answer is not the factually correct one. This often happens when the model is asked about niche topics for which it has limited training data, or when it is forced to generate a specific type of information (like a case citation) and "improvises" to fulfill the user's request.

The larger, frontier models like GPT-5 have greatly reduced the frequency of hallucinations for several reasons. First, their massive scale and improved training data mean they have a more robust internal "world model," making them less likely to generate nonsensical facts. Second, they have been heavily trained using techniques like Reinforcement Learning from Human Feedback (RLHF), where human reviewers explicitly penalize the model for fabricating information. Finally, when used in conjunction with grounded data

through techniques like Retrieval-Augmented Generation (RAG), where the model is forced to base its answer on a specific set of provided documents, the rate of hallucination can drop dramatically.

However, the risk never disappears entirely, particularly for older, smaller, or non-frontier models (like the open-source 'GPT-oss-20b' or many Llama variants). A lawyer using a less capable model to save costs must have their antennae up for potential fabrications.

4.2 Epistemic Triage: A Framework for Managing Risk

The key to using LLMs safely in a legal context is to adopt a mental framework of *epistemic triage*. As detailed by legal technologist Josh Kubicki, this means categorizing every query to an AI based on the risk associated with a potential hallucination.

- **Low Risk (Green Zone):** These are creative, synthetic, or brainstorming tasks where factual accuracy is not the primary goal. The cost of a hallucination is negligible. Examples include: "Brainstorm potential arguments for a motion to dismiss," or "Draft a first version of a client update email in a reassuring tone." You can trust the output of a frontier model for these tasks with minimal supervision.

- **Medium Risk (Yellow Zone):** These are tasks that involve factual information, but it is for background understanding or internal use, not for direct citation or dispositive reliance. An example might be: "Explain the core principles of the business judgment rule in Delaware corporate law." The output should be treated with professional skepticism and cross-referenced against your own knowledge,

but it doesn't necessarily require independent verification in a primary source database.

- **High Risk (Red Zone):** These are tasks that demand absolute factual accuracy. The output is a specific, verifiable fact that will be relied upon in a professional work product. Examples include: "What is the citation for *International Shoe Co. v. Washington*?" or "Provide the exact wording of 28 U.S.C. § 1331." For these queries, the output of an LLM must **always** be treated as an unverified lead. It must be independently and meticulously verified using an authoritative primary source like Westlaw, LexisNexis, or the official government source. To do otherwise is professional negligence.

Adopting this framework transforms the problem of hallucination from an existential threat into a manageable operational risk.

5 The Hybrid Workflow and the Strategic Imperative

The reality of legal practice in 2025 is that no single tool is sufficient. The so-far dreadful implementation of AI by traditional legal databases like Lexis and Westlaw means their tools are often cumbersome and lack the reasoning power of general-use models. Conversely, the hallucination risk of general-use models means they cannot be trusted for high-risk factual verification. The only responsible and effective path forward is a hybrid workflow.

For purely grounded answers, students and practitioners must learn to combine the best of both worlds. The workflow for researching a new legal issue should look something like this:

1. **Ideation and Exploration (General LLM):** Begin by using ChatGPT, Gemini, or Claude in "thinking mode" to explore the conceptual landscape of the issue. Use it to understand key doctrines, identify potential legal theories, and generate a list of relevant search terms and potential landmark cases.

2. **Lead Verification (Traditional Database):** Take the list of potential cases and statutes generated by the LLM and treat it as a set of unverified leads. Systematically look up every single case on Westlaw or Lexis. Read the actual opinions. Use KeyCite or Shepard's to ensure the cases are still good law.

3. **Synthesis and Drafting (General LLM):** Once you have a corpus of verified, reliable legal sources, you can return to the general LLM. Provide it with the full text of these verified cases and ask it to help you synthesize the arguments and construct the first draft of your brief or memorandum. This is a "Green Zone" task, as the model is now grounded in reliable data you have provided.

4. **Finalization and Citation (Human + Database):** The final step is the lawyer's alone. You must meticulously review the AI-assisted draft, refine the arguments with your own judgment, and use the tools within Westlaw or Lexis to ensure every citation is perfect.

5.1 The New Competitive Moat: Fine-Tuning and Proprietary Data

This leads to the final, and perhaps most important, strategic point. As access to powerful base models becomes a commodity, the source of durable competitive advantage for a law firm is no longer the AI model it uses, but

the proprietary data it uses to *customize* that model.

The process of fine-tuning—taking a pre-trained open-source model and continuing its training on a specific, high-quality dataset—has become significantly easier and more accessible. A law firm can take a model like 'DeepSeek-500B' and fine-tune it on its entire internal archive of past work: every brief, every contract, every memo, every client communication for the last twenty years.

The result is a new, proprietary model that is an expert in that specific firm's area of practice. It knows the firm's preferred drafting style, it understands the nuances of its key clients' businesses, and it can generate work product that reflects the accumulated wisdom and experience of the firm's partners. This fine-tuned model, running securely on the firm's private servers, becomes a strategic asset that cannot be replicated by competitors. A firm's data, once a passive archive for conflict checks, is now its most valuable raw material for building a defensible technological advantage.

6 The Next Horizon: The Rise of Agentic AI

The evolution of LLMs to this point has been revolutionary, but it has largely been confined to a transactional paradigm: a human issues a prompt, and the AI provides a single, self-contained response. The next great leap, which is already underway, is the shift from these single-turn "reasoning engines" to multi-step, autonomous "agents." An AI agent is a system that can take a high-level goal, independently break it down into a sequence of sub-tasks, execute those tasks using a variety of digital tools, and adapt its plan based on the results, all with minimal human intervention. This is the difference between asking an associate for a case citation and telling them to "prepare a draft of a motion to dismiss." This shift has profound

implications for the practice of law.

6.1 The State of the Art: Early Agents

The foundational technologies for agentic AI are now present in the frontier models. ChatGPT-5, for example, possesses a nascent ability to take a complex goal and formulate a multi-step plan. It can decide on its own to browse the web for information, write and execute code in a sandboxed environment to analyze data, and then synthesize the results into a final report. This is a crucial step beyond simple text generation; it is the ability to plan and execute a sequence of actions.

The experimental frontier of this technology can be seen in research previews like Google's Project Opal. Opal is not a chatbot in a window, but a proposed AI layer integrated into the operating system itself. It is designed to perceive what a user is doing on their screen, understand the context of their work, and proactively offer to orchestrate complex tasks across multiple applications. It might, for example, see a lawyer reviewing a contract in a word processor, and offer to cross-reference a specific clause with the firm's internal database of similar deals, all without being explicitly asked. While still in the research phase, projects like Opal show the clear direction of travel: away from AI as a tool you actively command, and toward AI as a persistent, context-aware assistant that helps manage your entire digital workflow.

6.2 A Glimpse into the Near Future: The Agentic Legal Workflow

While a fully autonomous AI legal associate is not yet a reality in August 2025, the component technologies are sufficiently mature that we can paint a plausible picture of a near-future workflow. Imagine a partner giving an AI agent the following high-level directive: "We've just been retained by Acme Corp in their patent dispute with Globex. Prepare a comprehensive memorandum analyzing the validity of Globex's 'widget' patent and outlining our initial defensive strategies."

The agent would then autonomously execute a workflow like the following:

1. **Issue Spotting and Planning:** The agent first parses the request and creates a detailed research and analysis plan. It identifies the core legal concepts (patent validity, novelty, non-obviousness) and formulates a series of sub-tasks: conduct a prior art search, analyze relevant case law on claim construction, research Globex's litigation history, etc.

2. **General Research:** The agent uses its web browsing capabilities to gather general background information on the technology in question, the companies involved, and the current state of the market.

3. **Specialized Research:** This is the critical step. The agent, using its "tool use" functionality, accesses specialized, third-party databases via their APIs. It logs into the firm's Westlaw account to retrieve relevant case law on patent validity in the relevant jurisdiction. It then queries a specialized patent analysis tool like Midpage or Harvey to conduct an exhaustive prior art search, looking for patents or publications that might invalidate Globex's claims. It pulls all this structured and

unstructured data back into its private workspace.

4. **Synthesis and Drafting:** The agent synthesizes the findings from all its research into a coherent, well-structured first-draft memorandum, complete with preliminary arguments and supporting citations.

5. **Adversarial Testing:** The partner then issues a new directive: "Now, adopt the persona of senior counsel for Globex. Read this draft memo and write a ruthless critique identifying every logical flaw, weak argument, and piece of contrary authority." A second AI instance (or the same agent in a different mode) "red teams" the work product, savaging its own initial draft.

6. **Rebuilding and Fortification:** The original agent takes this adversarial critique and uses it to rebuild the memorandum. It refines its arguments to preemptively counter the identified weaknesses, searches for additional supporting authority, and strengthens the overall logical structure of the document.

7. **Verification and Citation Check:** Finally, the agent performs a meticulous fact-checking and citation-checking process. It programmatically cross-references every factual assertion and case citation in its final draft against the verified source documents it retrieved from Westlaw and the patent office database, flagging any discrepancies for human review.

We are not quite there yet. The connections between tools can be brittle, and the model's ability to self-correct over long, complex tasks is still limited. Significant human oversight is required at each stage of the process I just described. But we are also not far away. Every individual component of that workflow is possible with the technology of today. The primary challenge is no longer one of capability, but of integration, reliability, and trust.

The gap between today's semi-autonomous assistants and tomorrow's fully agentic legal associates is closing at a startling pace.

7 Conclusion: The Expanding Universe

The world of Large Language Models in August 2025 is a far cry from the simple chatbots that captured the public imagination just a few years ago. It is an ecosystem of immense power, complexity, and strategic significance. We have moved from an era of simple prompting to one of conversational collaboration; from a reliance on a few proprietary models to a vibrant landscape of open-source alternatives; and from treating AI as a simple productivity tool to leveraging it as a strategic asset for creating proprietary knowledge systems.

The old critiques, while once valid, are increasingly aimed at the ghosts of past technologies. Yes, hallucinations remain a critical risk, but one that can be managed with disciplined workflows and a clear-eyed understanding of the technology's limitations. The path forward is not one of blind trust, nor of fearful Luddism. It is a path of critical engagement, continuous learning, and strategic adoption.

As you begin this course, your task is to discard your old assumptions. The universe of what is possible with this technology is expanding every day. Your challenge and your opportunity are to learn how to navigate it.