AI Under the Hood: What's in the Data Driving Your GenAI Tools and Why It Matters

The data that lies behind a generative artificial intelligence (GenAI) tool is just as important to consider as the user interface or the cost. Without trustworthy or relevant underlying information, the resulting AI-generated output will be less helpful or less trusted and result in inefficiencies as lawyers and staff work to fill the gaps in the GenAI's response.

BY KRISTOPHER TURNER

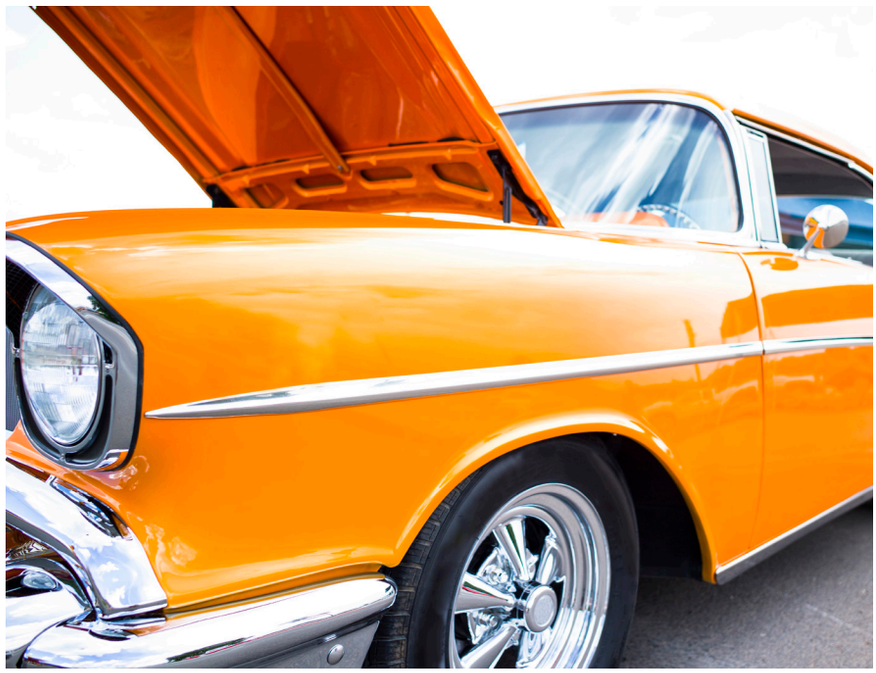
Everyone has a friend or family member who is *the* expert in one area of something. When you need a specific answer, you ask the person that obscure question because the person has spent considerable time absorbing and learning information that others simply do not know. Along the way, you probably have also learned not to immediately trust that person's knowledge about other subjects or to consider the source if you hear something that seems fantastic or outright wrong.

GenAI tools can act much like our expert-on-one-topic friends. Tools such as ChatGPT, Claude, Copilot, or Gemini are trained on sources and data that cover practically any topic that has been discussed or written about. However, it is not always clear which sources these general AI models use to become efficient and fast producers of information. Additionally, other models, such as Protégé (from Lexis+AI), CoCounsel (from Thomson Reuters), and Vincent (from VLex), are AI tools that are trained on specific data or information for a specific purpose. The sphere of their knowledge is purposefully limited to the law (and law-adjacent sources) so they can provide the user with a response based on answers that rely on legal research.

Knowing the difference between a general AI tool and one trained on specific sources can mean the difference between getting an accurate answer and becoming quickly frustrated with outcomes that either don't answer the question thoroughly or answer the question in a confused mixture of fact and fiction. While not always clear, the data that lies behind the GenAI tool is just as important to consider as the user interface or the cost. Without trustworthy or relevant underlying information, the resulting AI-generated output will be less helpful or less trusted and result in inefficiencies as lawyers and staff work to fill the gaps in the GenAI's response.

Kristopher Turner, U.W. 2020, is Associate Director of Public Services at the University of Wisconsin Law Library, Madison. Access the digital article at www.wisbar.org/wl.

kris.turner@wisc.edu



Taking AI to the Mechanic: Check the (Data) Engine

While GenAI tools like ChaptGPT have only been widely available since November 2022, much has changed in that short time. IBM provided a useful definition of GenAI tools in November 2023 that remains valid: "Generative AI refers to deep-learning models that can generate high-quality text, images, and other content based on the data they were trained on."[1] Understandably, many users focus on the generative-access aspect of these tools because they create images and text in seconds. However, the quality and breadth of this generated information is extremely dependent on the data on which it was trained.

When considering using a GenAI tool for legal work, lawyers must first consider the data that forms the pool of "knowledge" from which the tool can draw. There are times when this seemingly basic discovery can be opaque. Many general GenAI tools such as ChatGPT have trained on a vast quantity of resources, but pinning down exactly what is in that set of data can be difficult. Open AI, the company behind ChatGPT, explains that the tool is trained on data that is "freely and openly accessible on the internet" and OpenAI does "not intentionally gather data from sources known to be behind paywalls or the dark web."[2] Based on this, ChatGPT will train on sites such as Wikipedia, Reddit, publicly available social media and news outlets, YouTube, and freely available scholarship held in open-access digital repositories.

Additionally, Anthropic's Claude was at the center of a recent court case involving its use of books for training.[3] The federal judge on the case ruled that the use of books was not a copyright violation and fell under fair use. Anthropic, and other companies such as Meta and Apple, have also used scripts from popular TV shows such as *The Wire* and movies such as *The Godfather* to become more adept at communicating in a personable manner and to respond to pop-culture centric queries.[4] Other types of data included in these trainings are of particular interest to lawyers: court cases, statutes, court rules, and regulations, all of which are generally widely available.

It is obvious that general GenAIs have a vast amount of information from which to draw. This gives lawyers a roadmap of what to consider when interacting with a general tool such as ChatGPT, Claude, Perplexity, Copilot, or Gemini.

Hallucinations. Even newer models, with fine-tuning and more data, can hallucinate. Common sense suggests that if more data is provided, models will be able to reduce confusion and misleading answers. However, recent studies[5] have shown an increase in hallucinations from the newest AI models.[6] The reasons for this could be tweaks in the ways the tools provide or present information or poor information in the training data that poisons the final output. Even with advanced training and access to much of the world's knowledge, there remains the possibility that the GenAI tool will get something wrong.

Inappropriate information. Efforts are taken to weed out inappropriate, stereotypical, and hateful information, but that depends on both the AI and the humans working with it. In the face of myriad controversies, AI tools and the companies behind them have taken various steps to eliminate training data

that includes misinformation, spam, and hate speech. These steps can include human intervention, modification of training-data sources, or filters to keep out particular phrases. However, just as with cybersecurity measures, bad actors can improve their own methods to bring poisoned data into training, which results in misleading or offensive outputs. A GenAI tool that casts a wide net for its sources across the internet may ingest and train upon this type of information, undercutting the accuracy and trust of the tool.

AI training on AI-generated data. Much like a snake eating its own tail, some AI is already learning from datasets created by previous iterations, especially as so much of the generated information is posted freely online, where it becomes fair game for training (and perhaps goes undetected by the filters for hate speech because it is not offensive information, just bad information). Some research suggests that AIs trained on AI data will erode[7] and lead to "model collapse."[8] One especially striking example of AI information online is the prevalence of AI-generated misinformation news sources masquerading as official news.[9] If training data worsens, it inevitably follows that either the process of preparing the AI models will have to be modified or the outputs will also get worse.

Gaps. Be aware of what is *not* in the training data. As OpenAI notes, the company does not seek data behind paywalls. That leaves out a wealth of useful legal information, particularly secondary sources and other proprietary information that can be helpful in fundamental research and drafting. While there may be passing references to your favorite treatise, such as a State Bar of Wisconsin PINNACLE book on Books Unbound, the full text will not be part of the data that gives you a substantive answer. Some AI models may refer you to these resources (or to an attorney) if you want to learn more about a topic that is readily discoverable in a resource that is not part of the training for these general GenAI tools.

Specialized Mechanics: GenAI Tools with Curated Training Data

As helpful as a general tool like ChatGPT can be, newer AI models have become more specialized, which allows for a tighter focus on a single topic or subject. The future of GenAI may be coming into focus as individual AI agents serve as assistants on one aspect of work. A lawyer could have a "Personal Injury Bot" that only draws from data that focuses on cases and laws that relate to that area of practice along with templates for filings, appropriate calendars, and other relevant resources that a lawyer would refer to when doing this work manually. Established vendors have also created GenAI tools modeled in this vein by limiting the training data to trusted and verified primary and secondary sources.

The term for a large language model (LLM) that is grounded in more specific training data is retrieval augmented generation (RAG).[10] RAG is meant to focus the results of a GenAI output by more tightly controlling the source of its answer. This means that the AI tool is less likely to hallucinate because there are fewer chances for the incorrect information to be relied upon. It also creates a better sense of transparency because users can more easily determine the sources of an answer, either via direct reference to the controlling data by the GenAI tool or because the user is directly in control of the entire universe of the data from which a particular AI bot draws its training.

Specialized RAG tools are quickly multiplying in law practice. Protégé, CoCounsel, AI Assistant (from Bloomberg Law), and Vincent are all RAG tools, albeit RAG tools that still sit atop a large collection of resources. Each of these companies, and the content contained in their databases, are already trusted and commonly used by lawyers. The RAG tool takes this information and allows users to interact with it in a more dynamic manner, querying the sources more directly and producing answers to complicated questions complete with citations to the source of each of the RAG tool's assertions. Each legal database contains state and federal case law and statutes and also many peer-reviewed secondary sources written by lawyers and legal scholars and professors. This difference can dramatically alter how a GenAI tool answers legal research or drafting requests. A general AI model may draw upon legal blogs or the cases that are publicly available and at times cite them, but a legal research RAG tool will cite or quote those sources directly and better understand the language typically used in court filings because it has been trained on documents that use and define those terms.

An even more specialized type of RAG tool is becoming more prevalent. Lawyers and law firm staff can use an AI agent to query their own work product or to review their emails and calendaring. The goal is to maximize efficiency by offloading tedious tasks to AI while allowing lawyers and support staff to focus on areas that may require more human-oriented thinking, intervention, and input. A bot can be integrated into workflows and trained on a firm's archive of filings, client letters, and internal processes, which quickly get the bot up to speed on how the firm approaches various matters. The users can then interact with the bot to locate conflicts, draft situation-specific filings, and cross-check calendars and schedule meetings. A world in which attorneys and staff are freed from necessary but time-consuming routine tasks is one that AI may be able to provide as it trains on the appropriate data, giving it the chance to improve, become more accurate, and most important, answer questions in a format that matches the style and quality that would be expected of humans doing the same work.

Numerous GenAI companies offer the ability to create users' own "expert" AI. ChatGPT's GPTs shows off the capabilities and possibilities of these specifically curated and customized bots.[11] Other tools, such as Poe.AI, Google's NotebookLM, and Claude's Projects,[12] provide users the opportunity to add their own collection of sources to an AI and uncover patterns, arguments, and deeper understanding of the documents themselves. Other tools can be embedded in existing software, such as Microsoft's Copilot,[13] allowing users to search and create across all existing Microsoft Office documents.

When using such a tool, before enabling or uploading any proprietary or private documents, be sure to carefully read and consider the data-sharing and security agreements. A good rule of thumb, particularly for general AI tools and especially free tools, is to never upload or enter any

information that should not appear in a publicly available court filing.

Although RAG tools help minimize the frequency of hallucinations, hallucinations can still occur. GenAI tools infer relationships between concepts and words and those relationships can get lost when the tool focuses on the incorrect interpretation of how terms relate to each other. Complicated legal terminology is no exception to this issue, and the additional requirement of determining whether the assertion is still good law further underlines the necessity of a lawyer checking the output of the GenAI. AI can streamline workflows, but allowing it to create final products without human oversight and checking carries the risk that costly mistakes will occur.

Maintenance and Upkeep: Integrating Well-Trained AI Tools into Your Practice

The potential for these tools is obvious. AI-powered assistants that become experts in the area on which they have been trained can make work more efficient and, as the tools continue to improve, possibly more accurate. When deciding whether to purchase or integrate an AI bot into your practice, consider the following factors:

Will the AI scale? Practicing with a small sample of documents will likely provide proof of concept for how these tools can assist and enhance your work. However, scaling that coverage up to include your firm's entire digital archive may overwhelm either your staff or your chosen AI tool. Ensure that the bot can learn and grow along with your firm without becoming either too costly or technically insufficient to be worth the investment. Most of the major tools allow for scaling and integration, and practice with a tool should showcase the possibilities of upscaling.

At what level will you begin the integration? Consider training your AI bot on a single client's case files or only one attorney's work or a case that is already closed; the latter will allow you to query and ask the AI bot for output with which you are familiar. This closed universe of data will give you a stronger idea of the capabilities of the AI. It also allows the firm to slowly expand at a deliberate speed that matches need and limits costs.

How will you monitor the performance and encourage feedback? Staff training on what to expect from a RAG tool is crucial. If people expect the tool to answer a question about Taylor Swift even though the data underlying the bot is focused on the firm's work product, their expectations will go unmet, and their usage will not match the sunk cost. Encourage staff to give feedback on the functionality of the GenAI tool, which will then highlight deficiencies, which will then yield clues as to where further education or tweaking is needed, be it staff or AI training.

Conclusion

With the above considerations in mind, any law firm can begin exploring the addition of a cost-effective AI assistant to its legal technology infrastructure. Just like our friend or family member who is the self-proclaimed expert on one topic, a RAG tool can give focused and highly specific answers related to one area that someone needs to quickly understand. A firm that instead opts to purchase access to a general AI tool or legal research and drafting tool can also enhance their work by keeping in mind that the output is only as trustworthy as the data on which the GenAI tool is trained. As with a sports car, once you've peeked under the AI's hood, inspected the data powering the engine, and taken the wheel, you will have a more fundamental understanding of where these new tools can take you. **WL**

ENDNOTES

[1]Kim Martineau, *What Is Generative AI?* IBM (April 20, 2023), https://research.ibm.com/blog/what-is-generative-AI.

[2]OpenAI, *How ChatGPT and Our Foundation Models Are Developed*, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed (last visited Aug. 14, 2025).

[3]*See, e.g.,* Chloe Veltman, *In a First-of-its-kind Decision, an AI Company Wins a Copyright Infringement Lawsuit Brought by Authors*, NPR (June 25, 2025), https://www.npr.org/2025/06/25/nx-s1-5445242/federal-rules-in-ai-companys-favor-in-landmark-copyright-infringement-lawsuit-authors-bartz-graeber-wallace-johnson-anthropic.

[4]Alex Reisner, *There's No Longer Any Doubt That Hollywood Writing Is Powering AI*, The Atlantic (Nov. 2024), https://www.theatlantic.com/technology/archive/2024/11/opensubtitles-ai-set/680650/.

[5]Roland Moore-Colyer, *AI Hallucinates More Frequently as It Gets More Advanced – Is There Any Way to Stop It From Happening, and Should We Even Try?,* LiveScience (June 21, 2025), https://www.livescience.com/technology/artificial-intelligence/ai-hallucinates-more-frequently-as-it-gets-more-advanced-is-there-any-way-to-stop-it-from-happening-and-should-we-even-try.

[6]Jeremy Hsu, *AI Hallucinations Are Getting Worse – and They're Here to Stay*, NewScientist (May 9, 2025), https://www.newscientist.com/article/2479545-ai-hallucinations-are-getting-worse-and-theyre-here-to-stay/.

[7]*When AI's Output Is a Threat to AI Itself*, https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html. [This is behind a paywall.]

[8]Emily Wenger, *AI Models Collapse When Trained on Recursively Generated Data,* Nature (July 24, 2024), https://www.nature.com/articles/s41586-024-07566-y.

[9]*Tracking AI-Enabled Misinformation: 1,271 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools,* NewsGuard (May 5, 2025), https://www.newsguardtech.com/special-reports/ai-tracking-center/.

[10]Kim Martineau, *What is Retrieval-Augmented Generation?,* IBM (Aug. 22, 2023), https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

[11]*GPTs*, https://chatgpt.com/gpts (last visited Aug. 14, 2025).

[12]Poe, https://poe.com/explore?categorty=Official (last visited Aug. 14, 2025); NotebookLM, *Understand Anything*, https://notebooklm.google (last visited Aug. 14, 2025); Anthropic, *Collaborate with Claude on Projects* (June 25, 2024), https://www.anthropic.com/news/projects.

[13]*Microsoft 365 Copilot*, https://adoption.microsoft.com/en-us/copilot/ (last visited Aug. 14, 2025). **WL**