

THE RESEARCH UNIVERSITY PRESS

International
Computational Law Journal

VOLUME 2 / ISSUE 1 / CURATED SOURCE EDITION

An Introduction to Large Language Models for Ediscovery Professionals

John Tredennick · Dr. William Webber

MIT Computational Law Report · Originally published 2024-10-14

Independent reader edition under CC BY 4.0. This attribution cover was added; original source PDF pages follow unchanged. Original authorship and source identity are retained. MIT, Stanford University, Stanford Law School and CodeX have not sponsored or endorsed this republication.

Source, license and version evidence accompany the online article.

MIT Computational Law Report

An Introduction to Large Language Models for Ediscovery Professionals

John Tredennick Dr. William Webber

Published on: Oct 14, 2024

URL: https://law.mit.edu/pub/generativeaiforsmartpeople

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

Introduction

ChatGPT's release in November 2022 marked the beginning of a new era, the Generative AI era, that promises to radically transform and improve how we find, analyze and report on information across large volumes of unstructured data.

Not surprisingly, this new AI technology has become a focal point for discussion across almost every industry, including the legal profession. From large and small law offices to corporate legal departments and beyond, the potential for GenAI to reshape legal workflow is being explored, debated, and increasingly realized.

While the underlying technology is complex, our aim is to demystify GenAI, offering insights into its functionalities and uses that are both accessible and actionable for legal practitioners.

In Part One, we explore the fundamentals of GenAI's Large Language Models (LLMs), delving into key concepts such as their workings, training, and communication methods. We also address the ethical implications of using GenAI in legal practice, including data security and the potential for hallucinations.

In Part Two, we examine the practical applications of GenAI to legal workflow, with a particular focus on investigations and discovery, which is our primary area of expertise. We also discuss the growing use of GenAI to analyze and report on the large volumes of contracts that every large corporate legal department must manage.

Our hope is to show you that the capabilities we discuss are applicable (and beneficial) to almost every aspect of legal workflow. We then briefly turn to future developments in legal AI and offer guidance on how professionals can prepare for these changes. Throughout, our goal is to balance the promises of increased efficiency with a clear-eyed assessment of current limitations and risks.

Whether you're tech-savvy or just beginning to explore AI in your practice, we hope to equip you with the knowledge to navigate the GenAI revolution confidently. By embracing these technologies thoughtfully and ethically, legal professionals can enhance their practice, better serve clients, and shape the future of law.

Join us in exploring Generative AI for the legal profession. As your AI-powered assistant, it will unlock new levels of efficiency, insight, and innovation in your work, just as it has with ours.

Part One: What is Generative AI and How Does it Work?



Note: All images in this article (other than screenshots) were created by the authors using GPT 4o.

Generative AI is a type of artificial intelligence capable of creating new content, such as text, images, videos, audio, code, and other data. It achieves this by learning the patterns and structures from the training data it is exposed to and then using that knowledge to generate novel and original content with similar characteristics.

At the heart of Generative AI are advanced neural networks called Large Language Models (LLMs). These models are typically trained on vast amounts of data, comprising billions of text examples and other forms of media. Through this training process, GenAI models learn to recognize and replicate the underlying patterns and structures in the data, enabling them to create content that closely resembles human-generated work.

In Part One, we will delve deeper into the workings of LLMs, exploring key concepts such as training, context windows, and the relationship between ChatGPT and its companion LLM called GPT. We will also address important ethical considerations like data security and the potential for hallucinations in AI-generated content. Our hope is to give you a solid foundation in the fundamentals of Generative AI and LLMs, setting the stage for a later discussion of several practical applications that might whet your appetite for reimagining your own legal workflow.

What is GPT?

GPT stands for "Generative Pre-trained Transformer." It is a type of artificial neural network used in natural language processing tasks that uses deep learning techniques to generate human-like text. GPT models are

trained on vast amounts of diverse text data, allowing them to learn patterns and structures in natural language. Importantly, GPT models are "generative": that is, they are able to generate original text, based upon their learning. These models took the world by storm because of their ability to answer questions, create poems, analyze and summarize documents and carry on human-like conversations.

The term itself was coined by OpenAI, which gave a shortened version of that name to the first generative Large Language Model, which it called GPT. Not long after, the term generative pre-trained transformer became generic, used to describe these newly created Large Language Models that now represent the leading edge of artificial intelligence. All the current LLMs use variants of generative pre-trained architecture.

Different versions of these LLMs are referred to by numbers, e.g. GPT 3.0, GPT 3.5 or GPT 4.0. The more recent LLM versions come with added names and sometimes a reference to the size of their context window, e.g. GPT 4.0 Turbo (128k) or the somewhat confusingly called GPT-4o (128k). We will discuss the parenthetical numbers in a minute.

What are LLMs?

Because of the extensive training and expansive computational resources required for a GenAI model, they are often called Large Language Models ("LLMs"). GPT is an LLM but there are many others on the market today including Anthropic's Claude, Google's Gemini, Meta's Llama, Falcon and Mistral. Indeed, there are hundreds, if not thousands of generative pre-trained transformer models in existence today.



LLM models come in different sizes with different skills and intelligence levels.

Some LLMs are proprietary and available only from their publisher. Others are available through open-source licenses. In the latter case, you may need to host the model yourself, or use one that is available through a cloud provider like AWS. The proprietary models are SaaS based and typically accessed over the Internet through a secure API (Application Programming Interface).

While there is much to be said about the advantages and disadvantages of these competing forms of delivery, our focus is on using LLMs for discovery workflow rather than on which LLMs and which forms of delivery are better suited for your needs.

LLMs: Modern Supercomputers

LLMs require a massive amount of computing power and run on a large collection of expensive, specialized chips called GPUs or graphical processing units. One GPU chip suitable for LLM use may cost over $100,000.

The largest LLMs like GPT 4o or Claude 3.5 reportedly require tens of thousands of GPU chips to build. Microsoft reportedly used 25,000 NVIDIA chips for its GPT-4 implementation. An LLM once built may run on a single GPU, or small cluster of GPUs. Supporting multiple simultaneous users, however, will require many such compute nodes.

These expensive chips are needed because the prediction process that makes the LLM's output so valuable is mathematically intensive, requiring a huge amount of computing power. Efforts are underway to develop models that can run on smaller servers or even on a laptop. Indeed, Apple is reportedly working on models that can run, either in whole or in part, on your mobile phone. Building the models, however, will still require enormous computational capacity.

In addition to the hardware costs, there are significant operating expenses associated with running these chips daily. Although LLM providers don't publish exact cost figures, reports suggest that operating an LLM like GPT-4o can cost over a million dollars per day. Our suggestion? Don't plan on running one of the bigger LLMs at home, at least for the time being.

Ultimately, you can think of these LLMs as supercomputers, but with a level of depth, breadth, and power unprecedented in the history of computing.

Training an LLM

The primary goal for training an LLM is to enable the model to understand, generate, and reason with natural language in a way that is coherent, contextually relevant, and useful for a wide range of applications. That requires a two-stage training process.

First, the LLM must be trained on a massive amount of mostly Internet text including books, articles, websites and other textual sources. This is often called unsupervised learning as the neural network makes connections across the training examples and draws inferences it uses in creating responses. The process allows it to "understand" grammar, context, and a wide variety of topics.

We put the word "understand" in quotes because there is an ongoing debate about whether the LLM understands anything. Some critics call it a "stochastic parrot," arguing that while LLMs can produce content that appears coherent and contextually relevant, their output is essentially the result of statistically processing and regurgitating the vast amounts of data they have been trained on, without true understanding or consciousness.

In most cases, the training is supplemented by thousands of hours of human interaction focused on asking the model questions and providing feedback on its answers. This process of supervised training is referred to as "fine-tuning", and is critical to the model's fluency, effectiveness, and safety.

Training an LLM is expensive. The training cost for GPT-4 is estimated to be around $100 million, including the cloud computing costs of renting a 25,000 GPU cluster from providers like Microsoft. If cloud costs were $1 per A100 GPU hour, the cloud expenses alone would amount to around $60 million for a typical four-month training period.

Ultimately, the simple goal in training is to teach the model to predict what the next word in a sentence should be, based on the words that have come before. As crazy as that sounds, this is what LLMs do. They simply predict what the next words in a sentence should be based on the questions asked and the words that have gone before.

Training Cutoff

One crucial aspect to understand about LLM training is the concept of a "cutoff" or endpoint. During the training process, the model learns from vast amounts of data, continuously updating its parameters to better understand and generate language. However, once the training is complete, the model's parameters are fixed, and it can no longer learn from new data. (Of course, an LLM also cannot learn from data that it hasn’t seen, such as internal company documents or private emails, even if this data was generated before the training cutoff.)

This training cutoff has significant implications for the model's knowledge and capabilities. Everything the model has learned up until the cutoff point becomes its permanent knowledge base. It will not be able to adapt to or incorporate any changes, events, or new information that emerge after this point. For example, if an LLM's training data cutoff is January 2023, it will not have any knowledge of events or developments that occurred after that date.

It is essential for users to be aware of an LLM's training cutoff, as it directly influences the model's understanding of the world and its ability to provide up-to-date information. When interacting with an LLM, users should keep in mind that the model's knowledge is limited to the information available up until its training cutoff, and any queries or tasks related to post-cutoff events may yield outdated or inconsistent results.

It is also important to understand that an LLM has no memory. While it can communicate and provide responses to questions or prompts, it cannot remember your conversation once the answer is returned.

For that reason, we liken the trained model to a “brain in a jar” to reflect the fact that it has no memory and cannot learn from prompts or other information submitted to it. The LLM simply takes the information it is given and responds.

These limitations have important implications for LLM security. Since an LLM cannot use your prompt to broaden its knowledge base or remember previous conversations, it also cannot inadvertently pass prompt

information to other users. This means that the information you provide to an LLM remains secure and is not shared or learned by the model.

In summary, while LLMs are powerful tools for generating human-like text, their knowledge is limited by their training cutoff, and they do not have the ability to remember or learn from interactions. These characteristics, although they may seem like limitations, actually contribute to the security and privacy of the information shared with LLMs, as users can be assured that their data is not being stored or shared by the model.



LLMs are like a brain in a jar.

How Does it Carry on a Conversation?

That is the next obvious question. Many people have experienced carrying on a conversation with ChatGPT (or a competitor GenAI chatbot), and the initial experience can be eerie. ChatGPT seems to converse like a human, and some users engage in lengthy discussions with the software. If GPT, the LLM behind ChatGPT, has no memory, how can that happen?

The Role of ChatGPT

First, you need to grasp the difference between ChatGPT and GPT. ChatGPT is a software application designed to facilitate communication between users and GPT, the underlying Large Language Model (LLM) that analyzes and responds to questions. It provides a browser-based interface where users can enter their queries, known as "prompts."

When a prompt is submitted, ChatGPT sends it to GPT and then returns the generated answer to the user. The "Chat" part of ChatGPT saves the conversation history, allowing users to reopen previous discussions and continue the dialogue as if no time had passed. When a new request is made within an existing conversation, ChatGPT resends all the prior communications to GPT, enabling the LLM to maintain context and provide coherent responses throughout the interaction.

ChatGPT automatically saves your conversations, which means you can easily return to a previous discussion at any time. This feature allows you to pick up where you left off, even if you've had other conversations or taken a break in between. The saved conversations maintain the context of your earlier interactions, enabling GPT to provide coherent and contextually relevant responses when you revisit the dialogue. If you prefer not to

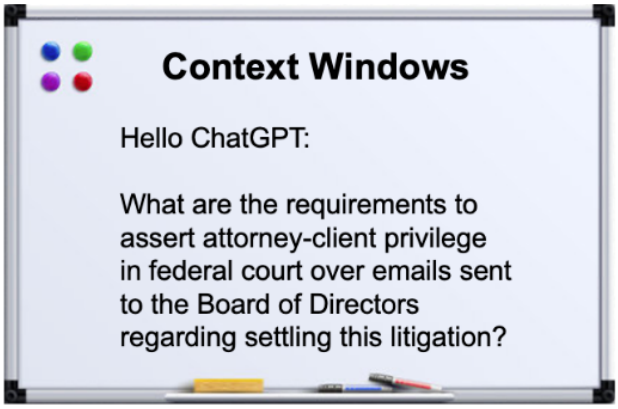
keep a record of your conversations, ChatGPT also offers the option to delete your chat history. This gives you control over your data and ensures that your discussions with GPT remain private if desired.

The Importance of Context Windows

GPT communicates with the separate Chat application through what is called a "context window." In our discussions, we liken it to a whiteboard, one that exists outside the "brain in a jar" but is accessible to GPT and ChatGPT.

ChatGPT starts a conversation by sending the text you enter to GPT via the context window. Put another way, it writes your prompt on the virtual "whiteboard."

GPT can read what is written on the whiteboard and write its answer back. The Chat application reads the answer and returns it to us by displaying it on the browser window. Once that answer is passed back to ChatGPT, it is erased, much like a computer's RAM memory is erased when you turn it off.



Context Windows are like a whiteboard. They are erased after the answer is returned.

Carrying on a Conversation

Now you understand how GPT can carry on an extended conversation even though the whiteboard is erased after each response. ChatGPT keeps track of your conversation and sends the earlier parts back to GPT each time you make a new request. GPT views the entire conversation (or as much of it as can fit on the whiteboard) and uses it to carry on the discussion. (Indeed, each request in the conversation may be handled by a different GPT computation node, drawn at random from the large bank of GPT nodes OpenAI maintains to service simultaneous users.)

Understanding the relationship between ChatGPT, GPT, and the context window is crucial to grasping how an LLM can carry on a conversation despite lacking inherent memory. Software applications like ChatGPT (and many others) keep track of your conversation and send it to GPT (or any other LLM) so that the LLM can provide relevant and coherent responses, creating the illusion of a continuous discussion.

(Note that recent versions of ChatGPT have added a “memory” feature, which distills certain facts about the user and their previous conversations, making them available across conversations. This “memory,” though, is still managed by ChatGPT, and the facts are passed to GPT through the context window. GPT itself is not “remembering” anything about you.)

Context Window Size Limitations

The most important thing to know about context windows is that the amount of text you can place on them (prompt plus answer plus conversation history) is limited.

When GPT 3.5 (the original engine for ChatGPT) was first released, the context window was 4,096 tokens, which translates to about 3,000 words. (Tokens include punctuation, and some words will be split into more than one token for technical reasons beyond the scope of this book). Thus, your conversation with GPT–including both questions and answers–was limited to the size of the context window. When your conversation grew larger than the window allowed, ChatGPT would cut out the first part of the conversation so you could continue to ask new questions. If the conversation continues beyond the size of the context window, GPT will begin forgetting aspects of your earlier conversation.

It's easy to imagine that a system which can only analyze 3,000 words of text would have practical limitations. You certainly couldn't ask it to read and comment on a book or even a lengthy article. You might ask GPT about a complex tax provision but certainly not about the tax code itself. Likewise, you couldn't and still can't ask GPT to read and analyze millions (or even thousands) of your discovery documents.

Thus, don’t confuse GPT or any Large Language Model with a search engine. Modern search engines can find information across millions of documents in milliseconds, but they can’t analyze the documents they have found. In contrast, GenAI engines can analyze documents you bring to them but only as many as can fit in the context window.

Enlargements in Context Window Sizes

In short order, LLM context windows have increased, moving from 4 to 8 to 16 and to 32k context windows. Recently, Anthropic (founded by people from OpenAI) released a 200k version of its LLM called Claude 3 and touted its ability to read the entirety of The Great Gatsby, not to mention the scripts for all nine versions of the Star Wars movies. This development generated excitement as OpenAI responded with GPT-4 Turbo (128k) and then the GPT-4o (also 128k) series. Anthropic reports that Claude 3 has the capacity to handle a one million token context window, though this is not generally available to users. Google similarly reports that its LLM, called Gemini, also has a one million token context window.

These were great advances from GPT's early days (literally just months before), but there are strong suggestions that increasing the context window to substantially larger sizes may not be feasible, either technically or due to cost considerations. Even if the windows can be made larger (which they undoubtedly will be), there is currently concern that the models cannot effectively remember everything read in large context windows. This may mean that they will overlook important details when giving their answers for inputs that use most of their nominal context window capacity, especially when performing more complex reasoning tasks.

At this point, all we can say is that the larger context windows open the door to using these powerful GenAI models for a variety of applications, including for investigations and discovery.

What about Hallucinations?

Many of us have heard the term “hallucinations" in the context of Generative AI but what are they? The concept is easy to understand but it has proven a bit unsettling for legal professionals to contemplate.



Hallucinations occur when the LLM makes up the answer to your question.

At its most basic, a hallucination is where an LLM confidently gives a detailed answer that seems plausible but is simply not true. Hallucinations typically arise because of the way LLMs are trained. LLMs are trained to predict the next word in a sequence based on the words that come before it. The LLM isn't concerned with whether the statement is true, only that it writes a convincing continuation of whatever text it is given.

As a result, LLMs can sometimes generate content that is fluent and plausible but not actually grounded in reality. The model may combine snippets of information from various sources in a way that seems coherent but is ultimately incorrect or fabricated.

A hallucination should be distinguished from a mistaken inference. A mistaken inference occurs when an intelligence misunderstands the import of a text, due to unjustified assumptions or faulty implications or the text’s own ambiguity. Humans make mistaken inferences: LLMs make them too. Hallucinations, on the other hand, are invented facts. A human producing the equivalent of an “hallucination” would be purely fabricating—deliberately inventing information they know to be untrue (or at least unfounded). An LLM, however, “sees” its task as the generation of plausible text, and is not always “conscious” that this generation is an invention. A key goal of the fine-tuning process LLMs are put through is to reduce the frequency of hallucinations, but fine-tuning alone does not entirely eliminate them.

A striking example of hallucination occurred when a lawyer cited fake cases generated by ChatGPT to a court. In the filing, the lawyer cited at least six cases that did not actually exist, with fake judicial decisions, bogus quotes, and bogus internal citations.[1]

The revelation came when opposing counsel couldn't find the cases and requested more information. Ultimately, the offending lawyer was called into court to explain the situation and faced the possibilities of sanctions (not awarded). However, since then there have been at least two other lawyers who made similar false filings and have received monetary sanctions.

These incidents highlighted the risks of lawyers relying on AI chatbots for legal research and writing without verifying the accuracy of the information provided. A simple antidote for this potential problem is to check key case citations before using them. Our lawyer author here received many memos from associates during his days as a litigation partner. He can’t remember a time he made a filing or appeared in court citing an important case he had not personally read. The lesson? Always check the source material whether you receive a memo from GPT or your crack associate.

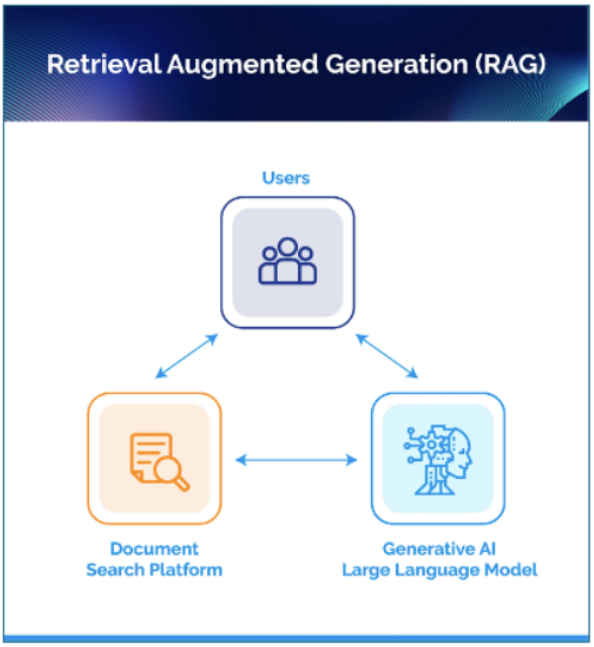
Reducing Chances for Hallucination

Hallucinations most often occur when an LLM is asked to answer a question based on its internal training. If the training data doesn't include the specific information requested, the LLM may sometimes generate a plausible-sounding but ultimately fabricated response to fill in the gaps. This is problematic, as it can lead to the dissemination of inaccurate or misleading information.

To address this issue, developers have explored alternative approaches to using LLMs, one of which is called RAG, or Retrieval Augmented Generation. RAG combines the power of the LLM's language understanding and generation capabilities with a more targeted and controlled input process.

In a typical RAG system, when the user asks a question, the first step is to search a collection of documents to find passages that are most relevant to answering it. Those passages are then inserted into the LLM's context window along with the user's original question.

The key difference between RAG and traditional LLM usage is that the LLM can be explicitly instructed to generate an answer based only on the provided passages, rather than relying on its general knowledge gained from training. By constraining the LLM to only the information present in the relevant passages, RAG helps to ensure that the generated answers are grounded in factual, verifiable information.



A typical RAG architecture.

Building on the concept of RAG, we have developed an even more stringent approach called Retrieval Exclusive Generation (REG). We coined the term 'Retrieval Exclusive Generation' to emphasize that this approach exclusively uses retrieved information, completely eliminating the LLM's reliance on its pre-trained knowledge for generating responses.

REG thus takes the principles of RAG a step further with the objective to provide safeguards against hallucinations if not eliminate them entirely..

In a REG system, we implement the following key strategies:

1. **Explicit instructions:** We instruct the LLM to base its analysis solely on the information provided through the context window, explicitly directing it not to rely on its own training or knowledge.
2. **Source citation:** The LLM is instructed to provide direct links to the source basis for all substantive statements in its response. This allows for immediate verification of the information provided.

This approach offers several advantages over traditional LLM usage and even standard RAG systems. By strictly limiting the LLM to the provided information and requiring source citations, REG significantly reduces the risk of hallucinations. It also provides a clear audit trail for the generated information, which is particularly crucial in legal contexts where accuracy and verifiability are paramount.

In Part Two, we will show you two examples of how we use a REG system to analyze discovery documents. In each example, the system first retrieves relevant documents from our database using AI-powered search techniques. Then, when prompted to analyze the retrieval results, the LLM is instructed twice to rely solely on the retrieved information sent to the Context Window.

In its response, the LLM provides its analysis of the documents or transcripts it has read followed by a direct link to the underlying source document or transcript section. This allows the legal professional to immediately verify the accuracy of the cited case and its interpretation, significantly reducing the risk of relying on hallucinated or misinterpreted information.

While REG systems can significantly reduce the occurrence of hallucinations, it's important to note that they are not infallible. Incorrect answers can still occur if the retrieved passages themselves contain inaccurate information. However, by employing strict anti-hallucination methodologies like those described in REG, the risk of hallucinations can be minimized, and any remaining inaccuracies can be more easily identified and verified.

As REG and similar techniques continue to evolve, we can expect to see further improvements in the accuracy and reliability of LLM-generated responses. By staying at the forefront of these developments and continuously refining our approaches, we can ensure that tools like DiscoveryPartner remain indispensable assets for legal professionals in their quest for truth and justice.

RAG/REG is also a way of overcoming the limitation of the training date cutoff and of making private information available to the LLM that was not contained in the training data—but more of this in Part Two.

Is the Data We Send to GPT Secure?

When ChatGPT was first released, many raised concerns about whether sending client information to the program might breach confidentiality obligations or risk a waiver of attorney-client or work-product privileges. For legal professionals at least, this is a serious question. Lawyers have ethical obligations to preserve client confidentiality, and they have a parallel duty to protect against an inadvertent waiver of attorney-client or work-product privileges.

Much of the concern arose because ChatGPT was originally released as a free public beta. As a condition of the free license, OpenAI reserved the right to use information contained in the prompts for testing and to improve later models.

Not surprisingly, legal professionals became concerned about the risk of sending confidential information to GPT, particularly about the possibility that the information might be disclosed later, either inadvertently by GPT or through its use in training new models. (In practice, user conversations are probably used in the fine-tuning stage, to hone the quality and form of GPT responses, in which case the danger of information leakage is reduced—but not to zero.)

In response, companies like OpenAI, Microsoft, and Anthropic began offering commercial licenses for their LLMs. These licenses include written promises that the information sent to the LLMs would not be used for any purpose other than your interaction with the LLM itself.

For example, Microsoft offers this statement regarding OpenAI, which it hosts on Azure:

> Your prompts (inputs) and completions (outputs), your embeddings, and your training data:
>
> - are NOT available to other customers.
> - are NOT available to OpenAI.
> - are NOT used to improve OpenAI models.
> - are NOT used to improve any Microsoft or 3rd party products or services.
> - are NOT used for automatically improving Azure OpenAI models for your use in your resource (The models are stateless, unless you explicitly fine-tune models with your training data).
> - Your fine-tuned Azure OpenAI models are available exclusively for your use.
>
> The Azure OpenAI Service is fully controlled by Microsoft; Microsoft hosts the OpenAI models in Microsoft's Azure environment and the Service does NOT interact with any services operated by OpenAI (e.g., ChatGPT, or the OpenAI API)."

OpenAI and Anthropic similarly promise not to review communications between the user (prompt) and the system's response. Here, for example, is the controlling provision from OpenAI's service agreement:

> "We do not use Content that you provide to or receive from our API ("API Content") to develop or improve our Services. We may use Content from Services other than our API ("Non-API Content") to help develop and improve our Services."

API stands for "Application Programming Interface." It is a software intermediary that allows two programs to talk to each other without human involvement. Commercial applications typically communicate with LLMs through a secure, encrypted API.

Can an LLM Share Confidential Information, Even by Accident?

From our earlier discussions, you know that the answer is no. The LLM itself cannot learn from the prompt information you send into its knowledge base, and it cannot remember your conversations. You also know that the LLM doesn't save the information it reads from the context window. Thus, the only risk here is that the LLM provider intercepts your communications and uses them for an illicit purpose.

In that regard, sending prompt information to a company like Microsoft is not materially different from storing files and email in Microsoft 365, Google Docs applications or storing files with other reputable companies. As a matter of fact, storing files with Microsoft or Google would seem to be far less secure than using an LLM for one simple reason: Office files and email typically sit on hosted servers for weeks, months or even years. Contrast that with the milliseconds that data sits in an LLM's context window before it is erased from the context window.

Reasonable Expectation of Privacy



More secure than Office 365? Yes, when you think about it.

While a legal discussion is beyond the scope of this book, courts have consistently held that using third party systems like these does not waive privilege so long as there is a "reasonable expectation of privacy."[2]

For example, courts have repeatedly taken the position that unencrypted email communications, even on a company server, do not result in a waiver of privilege so long as the person sending the communication had a "reasonable expectation of privacy." E.g., Twitter, Inc. v. Musk, C. A. 2022-0613-KSJM (Del. Ch. Sep. 13, 2022) (Musk used Tesla/SpaceX email servers for Twitter-related legal communications); Stengart v. Loving Care Agency, Inc., 990 A.2d 650 (2010) (personal legal communications made on work server). The courts did so notwithstanding the fact that Tesla and SpaceX explicitly reserved the right to inspect company emails for any purpose (including, presumably, abuse of the email privileges).

Likewise, the Standing Committee on Ethics and Professional Responsibility for the American Bar Association has repeatedly affirmed that email communications did not waive the privilege so long as the communicator had a "reasonable expectation of privacy" in the communication. ABA Comm. on Ethics & Prof'l Responsibility, Formal Op. 17-477 (leaving an open question about message boards and cell phone use).

The same undoubtedly holds true for the use of web hosting services like those offered by Microsoft, Google, or AWS as well as litigation support providers. All have access to your data at one point or another, but they are under contractual obligations not to exercise that privilege except to protect their systems from abuse or misuse.

Concerns about confidentiality and waiver are understandable with any new technology. Similar fears arose with cell phones, the internet, email, and cloud productivity suites. But so long as agreements and practices support a reasonable expectation of privacy, using LLMs via a commercial license should not create meaningful privilege risk, especially with the strong security measures in place for data sent to LLMs like GPT and Claude. The tremendous potential benefits should not be sacrificed for undue concerns.

So, are you risking a breach of confidentiality or privilege waiver by using an LLM through a commercial provider? Our answer is no, at least not if you are using a commercial license for the service. Providers like Microsoft, Anthropic, and other major large language model companies include solid non-disclosure and non-use provisions in their commercial contracts. They are easily as strong as the ones included in your Office 365 licenses. And, they provide the same reasonable expectation of privacy you have when you store email and office files in Azure or AWS.

Ethical Considerations of Using GenAI in a Legal Practice

While we've discussed data security and hallucinations, the ethical implications of using GenAI in legal practice merit further discussion. As smart legal professionals, we must grapple with a range of ethical considerations to ensure that our use of this powerful technology aligns with our professional responsibilities and the interests of justice.

The American Bar Association (ABA) and various state bar associations have begun to address these issues, providing guidance on how lawyers can ethically leverage GenAI while upholding their professional responsibilities.

ABA Formal Opinion 512: A Framework for Ethical GenAI Use

On July 29, 2024, the ABA Standing Committee on Ethics and Professional Responsibility issued Formal Opinion 512, offering the first comprehensive ethical guidance on lawyers' use of Generative AI[3]. This landmark opinion addresses several key areas that form the foundation of ethical GenAI use in legal practice:

1. Duty of Competence

The ABA emphasizes that while lawyers don't need to become GenAI experts, they must develop a reasonable understanding of the capabilities and limitations of the specific GenAI tools they use.[4] This includes:

- Engaging in ongoing education about GenAI through self-study, consultation with experts, or working with specialized vendors.
- Understanding potential risks and benefits, including the possibility of "hallucinations" or inaccurate outputs.
- Staying informed about the technology's evolving nature and its impact on legal practice.
- Recognizing when to seek assistance from technology experts or more experienced colleagues.

The duty of competence extends to adapting to new technologies, much as lawyers have had to adapt to ediscovery and digital case management in the past. As GenAI becomes more prevalent, this duty may evolve to include a basic understanding of AI principles and their application in legal contexts.

2. Confidentiality and Data Protection

Protecting client confidentiality is paramount when using GenAI. The ABA guidance emphasizes several key points:

- Lawyers must assess the risk of inadvertent disclosure of confidential information when inputting data into GenAI tools.[5]
- Informed client consent is required before entering confidential information into self-learning GenAI systems that may retain or learn from the data.[6]
- Law firms should implement clear policies and consult IT/cybersecurity experts to ensure GenAI tools adhere to stringent security and confidentiality protocols.[7]
- Lawyers should be aware of the data retention policies of GenAI providers and ensure they align with ethical obligations and client expectations.

The opinion draws parallels to existing guidance on cloud computing and third-party services, suggesting that using commercial GenAI tools with appropriate safeguards is generally permissible, similar to using services like Microsoft 365 or Google Workspace.

3. Independent Verification and Due Diligence

Given GenAI's potential for inaccuracies or "hallucinations," lawyers must independently verify any outputs used in client representation.[8] This point has been underscored by recent high-profile incidents where lawyers faced sanctions for citing non-existent cases generated by ChatGPT.[9]

In one striking example, lawyers Steven A. Schwartz and Peter LoDuca were each fined $5,000 and referred to grievance committees for potential further disciplinary action after submitting a legal brief containing fake cases generated by ChatGPT.[10] In another incident, attorney Jae Lee faced possible sanctions for citing a

nonexistent state court decision generated by ChatGPT, leading to a referral to a grievance panel by the U.S. Court of Appeals for the 2nd Circuit.[11]

These incidents highlight the critical importance of thorough verification and due diligence when using GenAI tools in legal practice. The ABA guidance suggests several best practices:

- Cross-referencing GenAI-generated information with established legal databases.
- Maintaining a healthy skepticism towards AI-generated content, especially novel or unexpected information.
- Documenting the verification process to demonstrate due diligence if questions arise.

4. Transparency and Client Communication

The ABA guidance[12] suggests that lawyers should consider disclosing their use of GenAI to clients, especially when it informs important decisions about representation.[13] This transparency may extend to:

- Discussing the use of GenAI tools in client engagement letters.
- Explaining how GenAI is being used to assist in the client's matter.
- Addressing any client concerns about AI use in their legal representation.

Furthermore, the ABA suggests that transparency may extend to disclosing GenAI use to courts and opposing counsel, particularly if there's any risk that the AI could have introduced errors or inaccuracies into the work. This aligns with the lawyer's duty of candor to the tribunal and fairness to opposing parties.

5. Supervisory Responsibilities

Managerial lawyers have additional ethical obligations when it comes to GenAI use within their organizations:

- Establish clear policies on GenAI use within their firms, including guidelines on appropriate use cases and security protocols.
- Provide comprehensive training for subordinate lawyers and staff on the ethical and practical considerations of using GenAI.[14]
- Ensure proper oversight of GenAI use by all personnel involved in client representation.
- Regularly review and update GenAI policies to keep pace with technological advancements and evolving ethical guidance.

6. Billing Practices

The ABA opinion addresses ethical considerations regarding billing practices when using GenAI tools:

- Lawyers cannot bill clients for time spent learning to use GenAI tools.
- Efficiencies gained through GenAI should benefit clients through reduced fees.
- Any use of GenAI that significantly impacts billing should be clearly communicated to clients.

- Firms should consider developing clear policies on how GenAI use is reflected in billing practices.

Additional Ethical Considerations

Beyond the ABA's guidance, legal professionals must grapple with several other ethical considerations as they integrate GenAI into their practice:

Bias and Fairness

LLMs, like any AI system, can perpetuate or amplify biases present in their training data. This presents particular challenges in legal contexts:

- Lawyers must be vigilant in identifying and mitigating such biases, particularly in areas like criminal justice or civil rights law.
- There's a need to critically evaluate GenAI outputs for potential discriminatory impact.
- Lawyers should consider using multiple AI tools and cross-checking results to mitigate individual system biases.
- The legal profession may need to develop standards for auditing AI tools for bias in legal applications.

Professional Judgment and Independence

While GenAI can be a powerful tool, it must not erode lawyers' professional judgment and independence:

- Lawyers must resist the temptation to over-rely on AI-generated content or analysis.
- The final legal judgment and responsibility must always rest with the human lawyer.
- There's a need to maintain and sharpen traditional legal skills alongside AI proficiency.
- Lawyers should view GenAI as a supplement to, not a replacement for, their expertise.

Unauthorized Practice of Law

As GenAI tools become more sophisticated, lawyers must be careful not to blur the lines between AI assistance and the unauthorized practice of law:

- Clear guidelines should be established to ensure that AI remains a tool, not a replacement for professional legal judgment.
- Lawyers must maintain control over the legal advice and services provided to clients.
- There may be a need for new regulations or guidelines defining the boundaries of permissible AI use in legal practice.

Intellectual Property Considerations

The use of GenAI in legal practice raises complex intellectual property questions:

- Lawyers must be aware of potential copyright issues when using AI-generated content.

- There may be questions about ownership and confidentiality of AI-generated work product.
- The legal profession may need to develop new norms around attribution and use of AI-generated legal content.

State Bar Guidance

Several state bar associations have also issued guidance on GenAI use in legal practice, often building upon or complementing the ABA's framework:

- The State Bar of California's Standing Committee on Professional Responsibility and Conduct issued practical guidance emphasizing confidentiality, compliance with AI-specific laws and regulations, and the importance of supervision and training.[15] Their guidance also stressed the need for lawyers to analyze relevant laws and regulations applicable to GenAI use, including AI-specific laws, privacy laws, and intellectual property considerations.
- The Florida Bar Board Review Committee on Professional Ethics issued Opinion 24-1, addressing confidentiality, oversight, legal fees, and lawyer advertising in the context of GenAI use.[16] This opinion provided specific guidance on how Florida lawyers should navigate the ethical complexities of AI integration in their practice.
- The New York City Bar Association's Formal Opinion 2024-5 provides general guidance on ethical considerations for using Generative AI tools in legal practice.[17] This opinion emphasized the need for lawyers to understand the limitations of AI tools and to maintain their independent professional judgment.
- The D.C. Bar issued Ethics Opinion 388 in April 2024, providing ethical guidance on the use of Generative AI in legal practice specifically for D.C. Bar members.[18]

These state-specific guidances often provide more detailed or nuanced advice tailored to the particular jurisdiction's ethical rules and legal landscape. Lawyers should consult both ABA and relevant state bar guidance when developing their approach to ethical GenAI use.

Charting New Ethical Territory

As we navigate these ethical considerations, we're charting new territory. The technology and our understanding of its ethical implications will continue to evolve. It's crucial that legal professionals engage in ongoing dialogue about these issues to ensure that we harness the power of GenAI to enhance our practice while upholding the highest ethical standards of our profession.

The integration of GenAI into legal practice is not just about adopting new technology; it's about reimagining the practice of law in the AI era. This requires a delicate balance between embracing innovation and maintaining the core ethical principles that have long guided the legal profession.

By staying informed, implementing safeguards, and maintaining a critical perspective on GenAI outputs, legal professionals can ethically leverage LLMs to improve efficiency and outcomes for their clients while fulfilling

their professional responsibilities. As the landscape continues to evolve, ongoing education, vigilance, and ethical reflection will be key to successfully navigating the AI-enhanced future of legal practice.

Part Two: Using Generative AI in a Legal Practice



In Part One, we explored the fundamentals of Generative AI and Large Language Models, gaining a deeper understanding of how these technologies work, the ethics of using them and their potential implications for the legal profession. With this foundation in place, we can now delve into a more practical side of our subject: Using GenAI and LLMs to reimagine and redefine legal processes and workflow.

In Part Two, we will explore several real-world examples that demonstrate how GenAI and LLMs can be used to enhance efficiency, accuracy, and cost-effectiveness in several legal workflow processes. Specifically, we'll dive into three key areas:

1. Using LLMs to find, analyze and report on key information in large document collections;
2. Using GenAI to analyze and report on large collections of transcripts or other lengthy documents such as chat messaging; and
3. Using GenAI to analyze and report on contracts.

Through these three examples, we hope to equip you with the knowledge and inspiration needed to start implementing these tools in your own work.

1. Using AI to Find, Analyze and Report on Information in Large Document Populations.

Imagine this scenario. You send the following assignment to your associate:

Bacardi Rum Dispute

> We represent Bacardi as a defendant in a trademark infringement dispute involving the French distiller Pernod and Cuba. The suit alleges that Bacardi is improperly using “Havana Club” as the name of a popular rum that it distills and distributes.
>
> The plaintiffs are claiming that Governor Jeb Bush improperly interfered with their claims in the Patent and Trademark court leading to its dismissal of the plaintiffs’ claims.
>
> We received a production of over 200,000 documents from the plaintiffs and need to see what evidence they have regarding Governor Jeb Bush. Please find all the relevant information you can from the production and give us a report on Governor Bush’s actions regarding this trademark infringement dispute.

With a traditional discovery workflow using standard discovery platforms, your associate might take these steps:

1. Create initial keyword searches to run against the document collection.
2. Review search hits and refine her searches based on the keywords and false hits returned.
3. Continue this process until she has reduced the population to a reasonable volume.
4. Review the search hits, weeding out the search misses until she narrows it down to the most relevant documents.

From there, the hard work begins. The associate must carefully read the relevant documents, determine what they say about the topic at hand and, ultimately, synthesize the information found into a coherent narrative. Finally, she must write up the memo answering the partner’s questions.

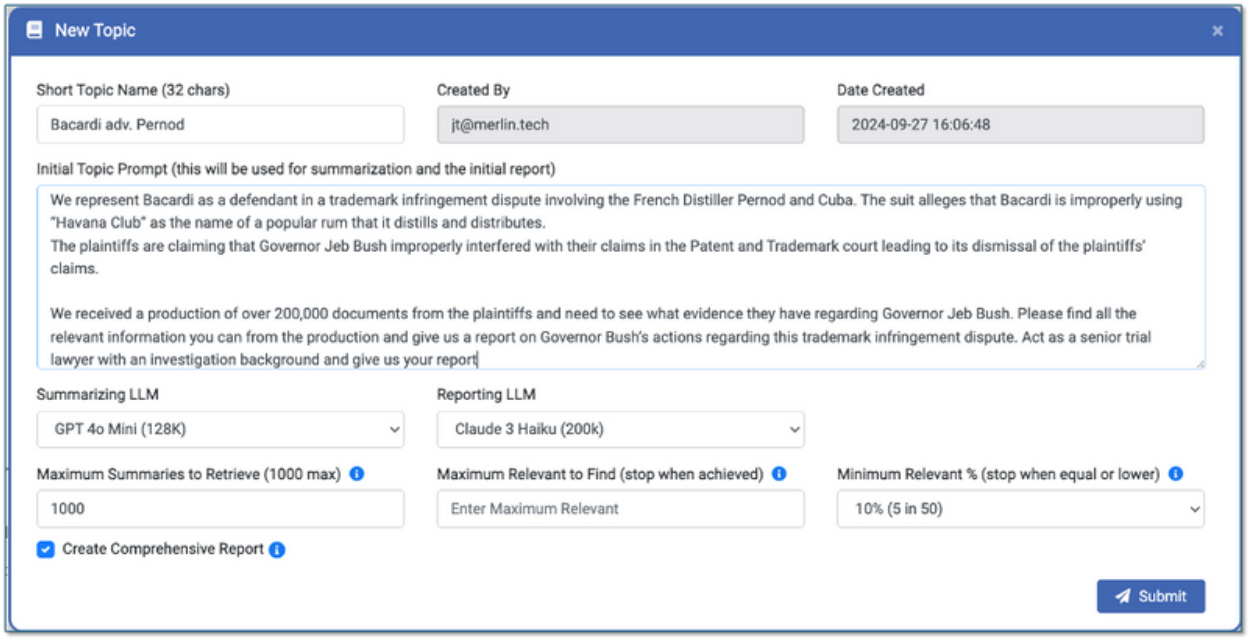
How long might this take? We will leave that to your imagination. But what if your associate could complete the assignment in minutes using GenAI and other machine learning tools to find, analyze and create your report?

Here is how it can happen with a GenAI powered document platform like DiscoveryPartner.

Sending an Initial Prompt

As noted, the production contains over 200,000 records. Only a relatively small number are about condominium associations in Florida.

To start the process, our associate filled out a simple “New Topic” form:



Topic Query

Here is how she described her Topic inquiry (aka Prompt).

> "We represent Bacardi as a defendant in a trademark infringement dispute involving the French Distiller Pernod and Cuba. The suit alleges that Bacardi is improperly using "Havana Club" as the name of a popular rum that it distills and distributes.
>
> The plaintiffs are claiming that Governor Jeb Bush improperly interfered with their claims in the Patent and Trademark court leading to its dismissal of the plaintiffs' claims.
>
> We received a production of over 200,000 documents from the plaintiffs and need to see what evidence they have regarding Governor Jeb Bush. Please find all the relevant information you can from the production and give us a report on Governor Bush's actions regarding this trademark infringement dispute. Act as a senior trial lawyer with an investigation background and give us your report."

LLM Choices

The system allows the user to choose between different LLM models for two phases of the workflow: 1) reading and summarizing documents retrieved from the search; and 2) reading and analyzing the relevant document summaries and preparing the report.

Why the choice? We have found that certain LLMs are quite good at summarizing documents. And, they cost about 30-50 times less than the larger models which are better at reporting. Making smart choices at this stage can save the user on LLM costs.

Summary Choices

The associate has several options regarding how deep she needs to go with this investigation:

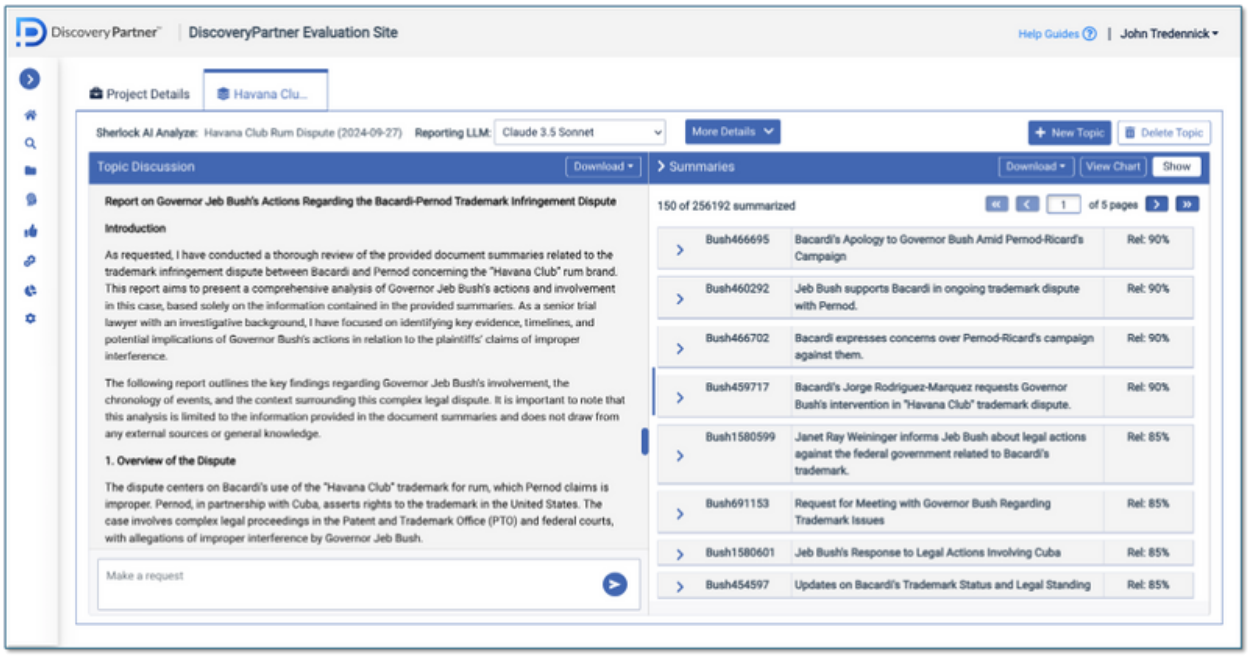
1. How many documents should the system find for summarization?
2. What is the maximum number of relevant documents the user should retrieve?
3. How low should the relevance rate go before the process should stop?

These options allow the associate to choose how deep in the collection she wants to go. She could set an absolute depth, e.g. the top 1,000 documents, or combine it with an order to stop when enough relevant documents have been found or the relevant rate of documents found goes below a set percentage, e.g. 10%.

In this case, our associate instructed the system to find and analyze up to 1,000 relevant documents, but asked the system to stop finding more documents when the rate of relevant documents found by several types of AI search used (NLP/Semantic, Algorithmic Keyword and a CAL classifier) dropped below 10%.

The Results

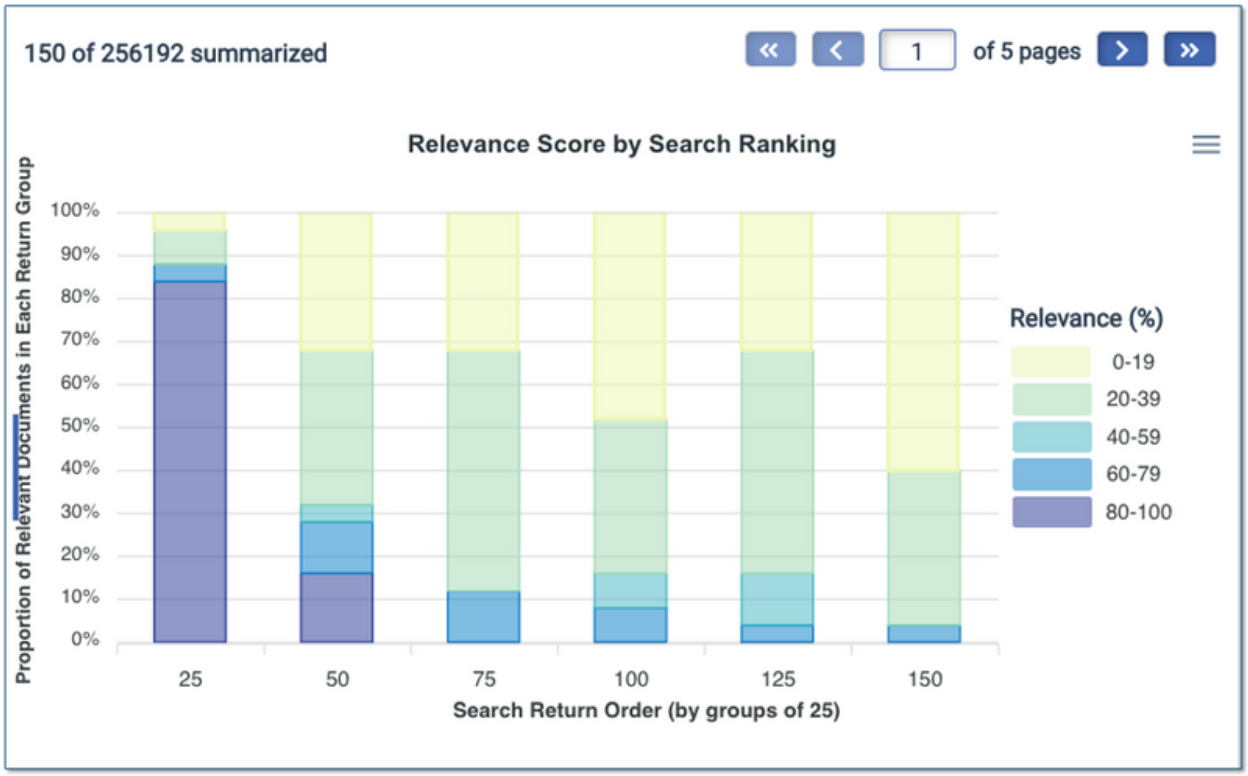
Here were the results.[19]



The entire process to find, analyze and report on the top 150 most relevant documents took 58 seconds–less than a minute from her starting click.

As you can see, our multi-AI system began by searching across 256,192 files but stopped retrieving documents after finding, analyzing and summarizing only 150 documents. Why was that? Because it ran out of relevant

files early on. You can see that in the “Relevance Score” chart the system created as an adjunct to its search.



We created this chart to help users understand the effectiveness of system searches. As you can see the first 25 documents found by the system were highly relevant, with the second group of 25 dropping in relevance but still at a reasonable level.

After that the system stopped finding relevant documents, primarily returning candidates with lower relevance scores (e.g. below 40%). Ultimately the system did its job. It stopped the search process because it wasn’t finding enough relevant information to justify continuing the process (and running up additional LLM costs).

The Report

Here is the report our associate received. It includes ID references to documents supporting the report’s conclusions.

Report on Governor Jeb Bush’s Actions Regarding the Bacardi-Pernod Trademark Infringement Dispute

Introduction

As requested, I have conducted a thorough review of the provided document summaries related to the trademark infringement dispute between Bacardi and Pernod concerning the “Havana Club” rum brand. This report aims to present a comprehensive analysis of Governor Jeb Bush’s actions and involvement in this case, based solely on the information contained in the provided

summaries. As a senior trial lawyer with an investigative background, I have focused on identifying key evidence, timelines, and potential implications of Governor Bush's actions in relation to the plaintiffs' claims of improper interference.

The following report outlines the key findings regarding Governor Jeb Bush's involvement, the chronology of events, and the context surrounding this complex legal dispute. It is important to note that this analysis is limited to the information provided in the document summaries and does not draw from any external sources or general knowledge.

1. Overview of the Dispute

The dispute centers on Bacardi's use of the "Havana Club" trademark for rum, which Pernod claims is improper. Pernod, in partnership with Cuba, asserts rights to the trademark in the United States. The case involves complex legal proceedings in the Patent and Trademark Office (PTO) and federal courts, with allegations of improper interference by Governor Jeb Bush.

2. Governor Bush's Involvement

2.1 Communication with Bacardi

Evidence suggests that Governor Bush maintained regular communication with Bacardi representatives, particularly Jorge Rodriguez-Marquez. These communications often involved discussions about the ongoing trademark dispute and requests for assistance from Bacardi (Bush459717, Bush460292, Bush466702).

2.2 Interactions with Patent and Trademark Office

Governor Bush appears to have engaged in communications with the PTO regarding Bacardi's trademark issues:

a) On May 23, 2002, Bush was advised to write to Jim Rogan, the newly appointed Director of the PTO, to advocate on behalf of Bacardi (Bush005324).

b) Bush sent a letter to James Rogan at the PTO supporting Bacardi's position, which was acknowledged by Bacardi as beneficial to their case (Bush007570).

c) The PTO later ruled on communications between Bush and PTO Director Rogan, stating that there were no legal or ethical issues regarding those communications (Bush657463).

2.3 Efforts to Expedite PTO Processes

Multiple documents indicate that Governor Bush was involved in efforts to urge the PTO to expedite its processes in favor of Bacardi:

a) Bacardi sought Bush's assistance in applying pressure on the PTO to expedite their cases (Bush460038, Bush460099).

b) Bush's office was reported to be working to resolve matters related to the trademark dispute (Bush460218).

3. Timeline of Key Events

- August 1997: U.S. federal courts resolved the matter in favor of Bacardi, confirming that Pernod-Castro had no rights to the Havana Club trademark in the U.S. (Bush466702).
- February 25, 2002: Bacardi met with the PTO, but reported no progress afterward (Bush459717).

- April-May 2002: Multiple communications between Bacardi and Governor Bush's office requesting assistance with PTO and OFAC issues (Bush459989, Bush460099, Bush005324).
- July 2002: Bush's office continued to work on resolving the trademark dispute (Bush460218).
- January 2003: The PTO ruled in favor of Bacardi, allowing them to expunge a trademark registration held by Fidel Castro (Bush657463).
- September 2002: Pernod Ricard filed a lawsuit against Governor Bush and the PTO concerning actions taken on behalf of Bacardi (Bush466967).

4. Allegations of Improper Interference

The plaintiffs claim that Governor Bush improperly interfered with their claims in the Patent and Trademark court, leading to the dismissal of their case. While the documents show Bush's involvement in supporting Bacardi's position, they do not provide direct evidence of improper interference. The PTO's ruling that there were no legal or ethical issues with Bush's communications suggests that his actions may have been within acceptable bounds (Bush657463).

5. Bacardi's Defense Strategy

Bacardi consistently maintained that the U.S. federal courts had already resolved the matter in their favor in 1997, and that Pernod-Castro had no rights to the "Havana Club" trademark in the U.S. (Bush466702, Bush466695). They viewed the ongoing challenges from Pernod as baseless and driven by publicity rather than legal merit.

6. Political Context

The trademark dispute appears to be intertwined with broader U.S.-Cuba relations and policies. Governor Bush's involvement may be seen in the context of his support for the Cuban-American community and U.S. policy towards Cuba (Bush581162, Bush542133).

Conclusion

While the documents demonstrate Governor Jeb Bush's active involvement in supporting Bacardi's position in the trademark dispute, including communications with the PTO and efforts to expedite processes, they do not provide conclusive evidence of improper interference as alleged by the plaintiffs. The PTO's ruling on the legality and ethics of Bush's communications suggests that his actions may have been within acceptable parameters. However, the extent of his involvement and its impact on the legal proceedings remain points of contention in this ongoing dispute.

We offer this first example to demonstrate the transformative potential of GenAI in legal workflows. What once might have taken days or even weeks of painstaking document review and analysis can now be accomplished in minutes. The AI-powered system not only rapidly identified relevant documents from a vast collection but also synthesized the information into a coherent, well-structured report. This level of efficiency allows legal professionals to focus their expertise on higher-level strategy and decision-making, rather than getting bogged down in time-consuming document review.

At the same time, it's important to note that while GenAI significantly accelerates the process, it doesn't replace the need for human expertise. The AI serves as a powerful assistant, providing a comprehensive starting point

for further analysis. Legal professionals still need to review the AI-generated report, cross-reference key documents, and apply their judgment and experience to the findings. The combination of AI efficiency and human expertise creates a powerful synergy that can dramatically enhance the quality and speed of legal work, ultimately benefiting both legal practitioners and their clients.

2. Using AI to Find, Analyze and Report Across Transcripts and Other Large Text Files.

LLMs can also be a game changer for reviewing and summarizing deposition and hearing transcripts. With a properly configured discovery platform, LLMs can not only summarize transcript testimony, but they can also answer questions across hundreds of transcripts and do so in seconds. Current transcript software can run keyword searches across transcripts, but it cannot answer your questions about witness testimony.

Deposition Summaries

The traditional approach for dealing with transcript testimony is to create a summary, typically in a Word or PDF format. These projects are often given to associates or senior legal assistants who read the transcript and summarize it as they go. One of the authors regularly created deposition summaries when he was a trial lawyer, dictating key points of the testimony along with page and line numbers.

A typical deposition summary might look something like this:

Leaking Barrels and Chemical Exposures at NECC	
36:23–37:1	Samples were taken during the filtering process and sent to the lab to ensure the product met the required quality before proceeding to the next stage.
37:2–37:11	After filtering, the reserve salt went through a drying process on a conveyor belt, similar to a potato chip dryer, and came out as a powder that was put into drums and shipped to customers for use as a metal stripper.
37:16–38:3	Mr. Nadeau recalls Metro-Atlantic also made other products like 40-S that went through a similar filtering process as reserve salt minus the drying, and products for waterproofing fabrics under the brand name Rane-Pel.
38:4–38:10	Various raw materials were used to make the waterproofing products, including fish oil, methanol, potash, waxes, caustic acid, and formaldehyde.
38:19–38:23	Mr. Nadeau acknowledges his memory may be somewhat contaminated from hearing other people's testimony, but most of it is his own memory.
39:1–40:3	Referring to his affidavit, Mr. Nadeau confirms that incoming barrels were stored on the ground at New England Container, some emitting chemical odors, and materials from the barrels leaked onto the soil.
40:23–41:19	The barrels were typically stacked on their sides in long rows, sometimes stacked on top of each other. NECC would receive hundreds of dirty barrels from various places for reconditioning.
42:2–42:7	Mr. Nadeau himself perceived and observed the leaks from the barrels, standing in the leaked material.

With an LLM, you can extend the process further, having the LLM include an overall summary of the deposition as we show here:

MERLIN

Deposition Summary

Joseph Nadeau

Case name: Emhart Industries, Inc. vs. New England Container Company, Inc., Providence Washington Insurance Company and Travelers Casualty & Surety Company, Consolidated with Emhart Industries, Inc. vs. United States Department of the Air Force, et al.

Date: Wednesday, June 12, 2013

Witness: Joseph Nadeau

Location: Offices of Duffy & Sweeney, LTD, 1800 Financial Plaza, Providence, Rhode Island

Overall Summary

In his deposition, Joseph Nadeau recounts his employment at New England Container Company (NECC) and Metro-Atlantic Chemical Company in the 1960s. He worked summers at NECC reconditioning drums that contained chemical residues, which often caused skin irritation and nosebleeds. Nadeau also worked briefly at Metro-Atlantic, where he was involved in manufacturing corrosive products like reserve salt using hazardous raw materials.

Nadeau describes the processes at both companies, including the use of dangerous chemicals, the generation of waste, and the disposal practices. At NECC, leaking barrels were stored on the ground, and workers were exposed to the leaks while handling them. The drum reconditioning process involved burning off residues, sandblasting, and steam cleaning. At Metro-Atlantic, the manufacturing processes generated black sludge waste that was partially hosed into floor drains leading to the canal. Leaks and spills from pipes and equipment were routinely washed down the drains into the "tailrace" as part of daily operations.

Nadeau also discusses the layout of the facilities, identifying various buildings and their functions using aerial photographs and maps. He mentions a small outbuilding called the Texas Tower at Metro-Atlantic, which he believes was used to produce a specialty product, possibly hexachlorophene, though he had little direct involvement with it.

Throughout his testimony, Nadeau emphasizes the pervasive chemical odors, the hazardous working conditions, and the potential environmental impact of the waste disposal practices at both NECC and Metro-Atlantic during his employment in the 1960s.

Deposition of Joseph Nadeau. You can view or download the full version of this summary (including transcript) here: https://bit.ly/3N5ZOgy.

We can then provide a hyperlinked table of contents with summaries and statements that distill key information from the deposition and provide instant links to the source testimony.

MERLIN

Chemical Raw Materials and Leaks at Metro-Atlantic ↻	
Raw materials arrived at Metro-Atlantic in various forms and were stored in above-ground tanks outside the building. The tanks were connected to the building via a network of pipes and valves, which allowed fluids to be transferred inside. Leaks and spills occurred from the pipes due to corrosion, and the spilled material was hosed into floor drains leading to the tailrace.	
69:3–69:17	Raw materials arrived at Metro for use in manufacturing chemicals via delivery truck or tanker in various packaging, including bulk pumped out of tankers, 55-gallon metal and fiber drums, and bags.
69:18–70:25	There were several above-ground storage tanks (three to six) at Metro on the outside wall over the wet area/tailrace, connected to the building via pipes. The tanks varied in size and some had large, heavy submarine tank covers that locked down with huge locking nuts.
71:1–71:20	Fluids were transferred from storage tanks to the Metro building via a pipe system with valves that allowed flow or transfer of tank contents to the desired destination, sometimes under pressure or gravity feed.
71:21–72:11	The storage tanks were connected to the Metro building by steel pipes with valves (quick valves and heavy turn valves) and flex pipes. The pipes allowed directing product from one point to another through a series of valves.
72:12–73:1	Leaks and spills occurred from the pipes, typically at unions, valves, or elbows, due to wear and corrosion from the harsh products like reserve salt. When leaks were noticed, the system would be shut down and a maintenance man would replace the leaking section.
73:2–73:18	The pipes eroded from the harshness of the contents like reserve salt, which would eat holes in tanks and mixing blades. Eroded pipes were observed where leaks occurred.
73:21–74:13	Leaks occurred inside the Metro building from pipes transporting raw materials from outside storage tanks. The leaks were noticeable when the product was moved under air pressure. Management was observant of costly leaks.

The hyperlinks go further, allowing the user to access the underlying transcript testimony with a click.

Deposition and hearing summaries are a standard way to extract information from their associated transcripts, but they are costly and time consuming to prepare. The summary excerpts you see above were created by an LLM in minutes, a fraction of the time it would take for a legal professional to do the job.

Beyond Summaries: Searching Across Transcripts

Once we mastered creating traditional deposition summaries, the next goal was to allow a user to search across multiple transcripts to get answers to questions about witness testimony. Our DiscoveryPartner system will allow users to create prompts that ask questions across hundreds of transcripts and get answers rather than search hits.

Based on the sections retrieved, the LLM will provide answers to questions about testimony during the deposition or hearing. Here is an example using two volumes of the Nadeau transcripts which provided the basis for the earlier deposition summaries.

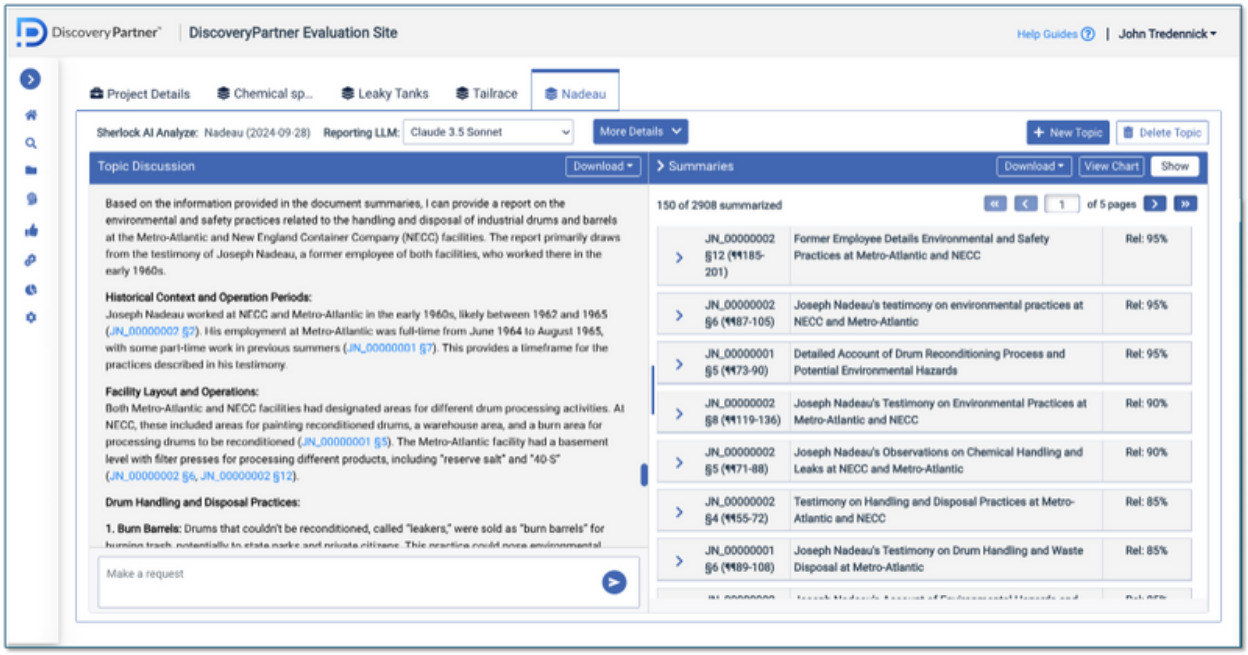
We start with a more complicated prompt which, interestingly, was created in part by an LLM.

> **Prompt:**
>
> I am investigating the environmental and safety practices related to the handling and disposal of industrial drums and barrels at the Metro-Atlantic and New England Container Company (NECC) facilities.
>
> I am interested in learning about tailrace and how it contributed to problems at the Metro superfund site. In your report, please answer these specific questions:
>
> 1. What is tailrace and how does it relate to superfund issues?
>
> 2. What causes tailrace and what is the resulting harm?
>
> 3. How do you mitigate problems from tailrace? Was there any testimony on this issue?
>
> 4. Who testified about tailrace and what did they say? Summarize in bullet points the testimony on this topic.
>
> I want to better understand the broader implications of finding tailrace at the Metro-Atlantic plant and how it relates to NECC's failure to comply with compliance with environmental regulations and the potential risks posed to public health and safety.

We ran this prompt across 150 deposition transcripts, most of which are about different subject areas to test the system's ability to find relevant information hidden in about 3,000 transcript sections. This is a true needle in the haystack exercise but against transcripts rather than documents.

The Report

In less than 50 seconds the LLM provided this report based on its review of 150 deposition transcripts (just under 3,000 sections).[20]



Here is the full text of the report:

Resulting Harm:

The resulting harm includes:

a) Contamination of soil and water (JN_00000002 §9, JN_00000001 §6).
b) Changes in water quality, including discoloration and murkiness (JN_00000001 §8).
c) Potential impacts on aquatic life, as evidenced by changes in visibility of fish and turtles (JN_00000001 §8).
d) Risks to public health and safety due to the release of hazardous materials into the environment (JN_00000002 §8, JN_00000002 §5).

3. How do you mitigate problems from tailrace? Was there any testimony on this issue?

The document summaries do not provide specific information on mitigation strategies for tailrace problems. There is no direct testimony on mitigation efforts. The focus of the testimonies is primarily on describing the practices that led to contamination rather than on remediation efforts.

4. Who testified about tailrace and what did they say? Summarize in bullet points the testimony on this topic.

Multiple unnamed witnesses provided testimony about tailrace and related environmental issues. Key points from their testimonies include:

- A witness described the tailrace as a discharge area for waste materials from the Metro-Atlantic facility (JN_00000002 §8).
- The same witness participated in hosing waste materials into drains that flowed to the tailrace, stating “It had no other place to go” (JN_00000002 §8).
- Another witness observed the tailrace changing colors and becoming murky over time (JN_00000001 §8).
- A witness testified about French drains in the basement that led to the outside and eventually to the river (JN_00000001 §6).
- Testimony indicated that vats used for storing and blending products were cleaned by draining contaminated water into French drains leading outside (JN_00000002 §12).
- A witness described seeing drums containing various substances stored on the ground, with materials leaking onto the soil around the plant (JN_00000002 §5).

Broader Implications

The broader implications of finding tailrace at the Metro-Atlantic plant and its relation to NECC’s failure to comply with environmental regulations include:

1. Widespread environmental contamination: The improper disposal practices led to the release of various hazardous materials into the environment, potentially affecting soil, water, and local ecosystems (JN_00000002 §8, JN_00000001 §8).
2. Public health risks: The discharge of industrial chemicals and waste products into the tailrace area raises significant health concerns for both workers and the surrounding community (JN_00000002 §8, JN_00000002 §5).
3. Regulatory violations: The practices described in the testimonies indicate a clear failure to comply with environmental regulations, potentially leading to legal consequences and designation as a superfund site (JN_00000002 §8, JN_00000002 §9).
4. Long-term environmental impact: The continuous discharge of contaminants over an extended period suggests that the environmental damage may be severe and long-lasting, requiring extensive cleanup efforts (JN_00000001 §8, JN_00000002 §12).

5. Worker safety issues: Testimonies reveal that workers were exposed to hazardous materials with inadequate protection, indicating a disregard for occupational safety standards (JN_00000002 §5, JN_00000002 §6).

Conclusion

In conclusion, the tailrace at the Metro-Atlantic facility played a significant role in the environmental contamination that led to its designation as a superfund site. The testimonies provided paint a picture of systematic disregard for proper waste disposal practices and environmental regulations, resulting in potentially severe and long-lasting impacts on the environment and public health.

Once again, this report was created in minutes, this time based on information quickly obtained across 150 transcripts.

Ultimately, legal professionals can best leverage this technology by using it as a first-line research tool, asking straightforward questions in natural language rather than crafting complex keyword queries.

Unlike traditional search methods that return a list of potentially relevant hits requiring manual review and synthesis, this GenAI-powered system provides documented, synthesized answers drawn from across multiple transcripts, dramatically reducing the time and effort needed to extract and compile relevant information for case preparation or strategic decision-making.

Perhaps most powerfully, this system can be invaluable during active depositions or trials. When unexpected issues arise, attorneys can quickly search across all previous testimony in seconds, even during short breaks, to better understand the context of an issue, identify contradictory statements, or prepare for cross-examination. This real-time access to comprehensive transcript analysis can significantly enhance an attorney's ability to adapt their strategy on the fly and effectively address emerging issues in the courtroom or deposition setting.

3. Re-engineering Contract Analysis with GenAI

Although this is a bit out of our wheelhouse, we would be remiss if we didn't mention that one of the most promising applications of Generative AI in legal practice is in the realm of contract analysis. This traditionally time-consuming task is ripe for AI-driven innovation, offering significant potential for improved efficiency and accuracy.

The Scope and Scale of Contract Analysis

The potential impact of GenAI on contract analysis is staggering when we consider the sheer volume of contracts in the business world:

- Large companies often manage thousands, if not tens of thousands, of contracts at any given time.
- Globally, the number of contracts that could benefit from AI-powered analysis likely stretches into the millions or even billions.

- These documents span a wide range, from vendor agreements and employment contracts to leases and licensing deals, each presenting its own complexities and nuances.

GenAI's capabilities extend across the entire contract lifecycle, revolutionizing every stage from creation to renewal.

At the outset, AI can generate initial drafts and customize templates, significantly streamlining the contract creation process. During review, GenAI tools can rapidly analyze lengthy contracts, extracting key information and summarizing important clauses. This ability to quickly distill the essence of complex legal documents allows lawyers to focus their expertise on more nuanced aspects of agreements.

Key Applications and Benefits in Contract Analysis

While we can't cover all the ways GenAI can improve contract management, here are several key applications and their benefits:

1. Effective Contract Review

GenAI can quickly analyze lengthy contracts, extracting key information and summarizing important clauses, while also flagging potential risks or anomalies. For instance, an AI could identify:

- Non-standard clauses or deviations from company templates
- Missing essential provisions
- Inconsistencies within the document
- Potential conflicts with local laws or regulations

This automated first pass can significantly reduce the time lawyers spend on initial contract review, allowing them to focus their expertise on addressing the flagged issues.

Benefits**:** Enhanced efficiency, improved accuracy, and early risk identification.

2. Drafting Assistance

While not replacing legal expertise, GenAI can assist in drafting by suggesting language for standard clauses, adapting existing clauses to new contexts, and ensuring consistency in terminology.

Benefits: Time savings, reduced human error, and enhanced risk management.

3. Data Extraction and Comparison

GenAI facilitates efficient extraction of key contract data, such as parties involved, important dates, and financial terms. It can also extract specific clauses from multiple contracts for easy comparison. This is invaluable when:

- Reviewing a large volume of contracts for due diligence
- Comparing new contracts against a company's preferred language
- Analyzing how certain clauses have evolved over time across different agreements

These capabilities can also be useful when migrating contracts to new management systems or during mergers and acquisitions.

Benefits: Streamlined due diligence processes, improved standardization, and more informed decision-making.

4. Language Translation

For international contracts, GenAI can provide rapid, accurate translations, helping lawyers who may not be fluent in the contract's original language to understand its contents quickly.

Benefits**:** The ability to have a working understanding of documents written in a different language with reduced risk of confusion and misunderstandings.

These kinds of capabilities will lead to significant cost reductions for law firms and legal departments while providing better contract management services for their clients. By automating time-consuming tasks, GenAI allows legal professionals to handle larger volumes of contracts more quickly and focus on higher-value work that requires human judgment and creativity.

Generative AI is rapidly redefining contract analysis for corporations and law firms, offering unprecedented efficiency and insights. As the technology matures, it will be crucial for legal professionals to adapt their skills and workflows to effectively leverage these tools. Those who successfully integrate GenAI into their contract analysis processes will be well-positioned to provide enhanced value to their clients and organizations, staying competitive in an increasingly tech-driven legal landscape.

4. The Future of GenAI in Legal Practice

As we conclude our exploration of Generative AI in legal practice, it's crucial to cast our gaze forward. The rapid advancement of GenAI technologies promises to reshape the legal landscape in ways we're only beginning to imagine. Smart legal professionals should not only adapt to these changes but position themselves at the forefront of this technological revolution.

Here are some of the developments we expect to see in the not too distant future.

Enhanced Natural Language Processing

Future iterations of LLMs will likely demonstrate an even more nuanced understanding of legal language and context. This could lead to AI systems capable of drafting complex legal documents with minimal human intervention, or even engaging in preliminary legal analysis of intricate cases.

Multimodal AI

While current LLMs primarily work with text, future systems are and will integrate audio, visual, and even tactile data. Imagine an AI that can analyze courtroom videos, scrutinize evidence photographs, or process audio recordings of witness testimonies, providing a comprehensive analysis across multiple data types.

Automated Case Law Analysis

We may see AI systems that can autonomously review vast repositories of case law, identifying relevant precedents and legal trends with unprecedented speed and accuracy. This could dramatically streamline legal research and case preparation processes.

AI-Assisted Decision Making

While AI will not replace judicial decision-making, we might see systems that assist judges by providing comprehensive case summaries, relevant precedents, and potential implications of various rulings.

Preparing for the Future

As these developments unfold, legal professionals must take proactive steps to stay relevant and leverage these new technologies effectively:

1. **Continuous Learning:** Make ongoing AI education a priority. This doesn't mean becoming a data scientist, but rather staying informed about AI capabilities, limitations, and ethical considerations in legal contexts.
2. **Ethical Leadership:** As AI becomes more prevalent, legal professionals will need to take the lead in shaping ethical guidelines and best practices for AI use in law. Engage with bar associations and legal tech communities to contribute to these important discussions.
3. **Interdisciplinary Collaboration:** Build relationships with technologists and data scientists. The future of law will likely involve close collaboration between legal and tech professionals.
4. **Become AI Literate:** While you don't need to code, understanding the basics of how AI systems work will be crucial. This knowledge will help you better leverage AI tools and identify their limitations.
5. **Improve Human Skills:** As AI takes over more routine tasks, focus on developing skills that AI can't easily replicate – critical thinking, emotional intelligence, creative problem-solving, and nuanced communication.
6. **Embrace Change:** Cultivate a mindset that sees technological change as an opportunity rather than a threat. The most successful legal professionals will be those who can adapt quickly and creatively to new technological landscapes.

The integration of GenAI into legal practice is not a distant possibility–it's today’s reality. By staying informed, adaptable, and ethically grounded, smart legal professionals can harness the power of these technologies to enhance their practice, better serve their clients, and contribute to the evolution of the legal profession. The future of law is being written now, and with the right preparation, you can play a pivotal role in shaping it.

Final Thoughts

Throughout this article, we have explored the fundamentals of Generative AI and Large Language Models, delving into their inner workings, capabilities, and limitations. Our primary goal was to provide smart legal professionals with the knowledge and tools necessary to harness the power of GenAI and LLMs in their investigation and discovery practices.

By focusing on several discovery workflow examples, we sought to demonstrate the transformative potential of GenAI in streamlining and enhancing critical tasks such as document review, analysis, and transcript review. From automating the classification and summarization of documents to extracting key insights from vast amounts of data, the integration of LLMs like GPT into discovery workflows marks a significant advancement for the profession.

As we have seen, LLMs can dramatically improve the efficiency and accuracy of these tasks, enabling legal teams to quickly identify relevant information and make better use of their time and resources. This, in turn, allows legal professionals to devote more attention to high-value activities such as developing trial and settlement strategies, exercising professional judgment, and providing sound advice to clients.

The promise of GenAI extends far beyond simply making existing processes more efficient. This technology has the potential to fundamentally reshape the very nature of legal work, opening new possibilities and redefining the contours of the profession. As we stand at the threshold of this new era, it is up to smart legal professionals to seize the opportunities presented by GenAI.

Appendix: GenAI Terms

Here are several terms smart people should know about Generative AI. These concepts are at the heart of this new form of artificial intelligence and will help you better understand our subject.

- **API (Application Programming Interface):** A set of protocols, routines, and tools for building software applications. In the context of GenAI, APIs allow users to interact with and access the capabilities of LLMs through a defined set of commands and inputs.
- **Generative AI (GenAI):** A type of artificial intelligence that can generate new content, whether it's text, images, music, or other forms of media, based on its training and the input it receives. This is accomplished through machine learning models that have been trained on large datasets, enabling them to recognize patterns, styles, or structures in the data.
- **GPT (Generative Pretrained Transformer):** A form of GenAI designed to understand, process, and generate human-like text based on the input it receives. As a legal professional, think of it as an advanced legal assistant or associate that can help with some pretty complex reading, analyzing, and writing tasks.

- **ChatGPT:** A web-based application offered by OpenAI that allows users to interact with GPT (i.e., send information through prompts) and receive answers. It runs on GPT but is not the same as GPT. Think of it as a front-end gateway, but not the only gateway to GPT.
- **Context Window:** The amount of text an LLM can process and consider at one time, which affects its ability to maintain coherence and relevance in longer conversations or documents.
- **Hallucination:** An AI phenomenon where the model generates plausible but factually incorrect information, which is particularly crucial for legal professionals to be aware of when using GenAI.
- **Large Language Model (LLM):** GenAI systems (often called models) like GPT, Claude, Gemini, Llama, and now hundreds of others that are specifically designed to understand, generate, and interact with human language. These models are "large" both in terms of the size of their neural network architecture (the complex web of interconnected nodes that process and store information) and the volume of data they have been trained on.
- **Neural Network Architecture:** The structure and organization of an artificial neural network, which consists of interconnected nodes (neurons) arranged in layers. This architecture allows the network to learn and process information by adjusting the strength of the connections between nodes based on the input data and desired output.
- **Prompt:** The initial input or instruction given to the GenAI model to elicit a specific response or output. Prompts can range from simple questions, commands, or statements to more complex scenarios or instructions, depending on the desired outcome. For example, a prompt could be "Write a summary of the key arguments in the Smith v. Johnson case."
- **Reinforcement Learning:** A type of machine learning where the AI model learns to make decisions or take actions based on feedback in the form of rewards or penalties. In the context of LLMs, reinforcement learning involves human interaction, such as asking the model questions and providing feedback on its answers to improve its performance.
- **Token:** A unit of data sent to or received from an LLM during the course of performing its services. A token may be a word, part of a word, punctuation, or a mix of the above and is on average approximately four characters in length. A rough guide is that 750 words equate to about 1,000 tokens. The number of tokens an LLM can process in one go is often referred to as "context length" or "context window."
- **Unsupervised Learning:** A type of machine learning where the AI model learns to identify patterns and structures in data without explicit guidance or labeled examples. In the context of LLMs, unsupervised learning involves training the model on vast amounts of text data, allowing it to learn language patterns and relationships on its own.

About the Authors



John Tredennick (JT@Merlin.Tech) is the CEO and founder of Merlin Search Technologies, a software company leveraging generative AI and cloud technologies to make investigation and discovery workflow faster, easier, and less expensive. Prior to founding Merlin, Tredennick had a distinguished career as a trial lawyer and litigation partner at a national law firm.

With his expertise in legal technology, he founded Catalyst in 2000, an international ediscovery technology company that was acquired in 2019 by a large public company. Tredennick regularly speaks and writes on legal technology and AI topics and has authored eight books and dozens of articles. He has also served as Chair of the ABA's Law Practice Management Section.



Dr. William Webber (wwebber@Merlin.Tech) is the Chief Data Scientist of Merlin Search Technologies. He completed his PhD in Measurement in Information Retrieval Evaluation at the University of Melbourne under Professors Alistair Moffat and Justin Zobel, and his post-doctoral research at the E-Discovery Lab of the University of Maryland under Professor Doug Oard.

With over 30 peer-reviewed scientific publications in the areas of information retrieval, statistical evaluation, and machine learning, he is a world expert in AI and statistical measurement for information retrieval and ediscovery. He has almost a decade of industry experience as a consulting data scientist to ediscovery software vendors, service providers, and law firms.

Footnotes

1. Lawyer cites fake cases generated by ChatGPT in legal brief, found at https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/. See also, AI is creating fake legal cases and making its way into real courtrooms, with disastrous results, found at https://theconversation.com/ai-is-creating-fake-legal-cases-and-making-its-way-into-real-courtrooms-with-disastrous-results-225080. ↩
2. We have written an article and given several webinars on the subject, one for U.S. audiences (with Professor William Hamilton of the University of Florida Law School) and one for Londoners (joined by Thomas Leyland of Dentons). You can find a copy of our August 2023 article Are LLMs Like GPT Secure? Or Do I Risk Waiving Attorney-Client or Work-Product Privileges?, found at https://merlin.tech/llm-security/ (Law 360, 8/17/23). You can watch both the U.S. and London webinars here: https://merlin.tech/headlines/programs-webinars/ ↩

3. American Bar Association, Formal Opinion 512 (2024), foiund at https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf. ↩
4. Breaking Down the ABA's Guidance on Using Generative AI in Legal Practice, found at https://www.2civility.org/breaking-down-the-abas-guidance-on-using-Generative-ai-in-legal-practice/ hereafter "Civility". ↩
5. See Civility at note 4. ↩
6. See Civility at note 4. ↩
7. See Civility at note 4. ↩
8. See Civility at note 4. ↩
9. AI is creating fake legal cases and making its way into real courtrooms, with disastrous results, found at https://theconversation.com/ai-is-creating-fake-legal-cases-and-making-its-way-into-real-courtrooms-with-disastrous-results-225080. ↩
10. Lawyer cites fake cases generated by ChatGPT in legal brief, found at https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557. ↩
11. NY Lawyer Faces Possible Sanctions for Citing Phony ChatGPT Case, found at https://news.bloomberglaw.com/business-and-practice/ny-lawyer-faces-possible-sanctions-for-citing-phony-chatgpt-case. ↩
12. American Bar Association, Formal Opinion 512 (2024) found at https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf. ↩
13. See Civility at note 2. ↩
14. See Civility at note 2. ↩
15. State Bar of California, Standing Committee on Professional Responsibility and Conduct, Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law (Nov. 16, 2023), found at https://www.calbar.ca.gov/Portals/0/documents/ethics/Generative-AI-Practical-Guidance.pdf. . ↩
16. Florida Bar Board Review Committee on Professional Ethics, Opinion 24-1 (2024), found at https://www.floridabar.org/etopinions/opinion-24-1/. . ↩

17. New York City Bar Association, Formal Opinion 2024-5: Ethical Obligations of Lawyers and Law Firms Relating to the Use of Generative Artificial Intelligence in the Practice of Law (2024), found at https://www.nycbar.org/reports/formal-opinion-2024-5-generative-ai-in-the-practice-of-law/#:~:text=OPINION%3A%20When%20using%20generative%20artificial,the%20duty%20to%20supervise%20both. ↩

18. D.C. Bar Ethics Opinion 388 (April 2024), found at https://www.dcbar.org/for-lawyers/legal-ethics/ethics-opinions-210-present/ethics-opinion-388. ↩

19. You can see a video showing the creation of this report here: https://bit.ly/4dpDGss. ↩

20. You can view a video of the creation of this report here: https://bit.ly/4enpDoi. ↩