MERLIN ®

GenAI for Smart People

An Introduction to Large Language Models for Smart Ediscovery Professionals

By John Tredennick and Dr. William Webber

It's been more than a year since ChatGPT upended our thinking about artificial intelligence. Over that time, discussions about Large Language Models like GPT (the engine behind ChatGPT) have taken center stage in just about every field of endeavor, but especially in the legal profession. At the recent LegalWeek conference, nearly every vendor on the exhibit floor touted their new Generative AI integrated software–or at least talked about how GenAI would be added to their product "real soon."

This Guide provides an introduction to Generative AI for smart legal professionals. While someone will no doubt release a "GenAI book for Dummies" in the near future, we don't see dummies in our profession. Rather, it is filled at all levels with smart people, many of whom would like to better understand how GenAI can make their practices more efficient and effective. We hope you enjoy it.

MERLiN

Executive Summary*

"**GenAI for Smart People**," authored by John Tredennick and Dr. William Webber, serves as an essential guide for legal professionals navigating the integration of Generative AI (GenAI) into ediscovery workflows. The article breaks down complex concepts into digestible insights, offering a forward-looking perspective on the technology's transformative impact.

Introduction to GenAI: The guide introduces GenAI, emphasizing its pivotal role in enhancing legal practices. It sheds light on how GenAI, particularly through Large Language Models (LLMs) like GPT, has become central to discussions in the legal field, promising to streamline processes and improve efficiency.

Role of LLMs: The authors elaborate on LLMs' functionality, focusing on their ability to process and generate text based on extensive training. This section underscores the computational prowess and the intricate technology driving these models, highlighting their significance in text-based applications.

Training and Operational Mechanics: The paper details the rigorous training LLMs undergo, including reinforcement learning and the implications of a fixed knowledge base post-training cutoff. It brings attention to the models' limitations and the ongoing debate around their ability to truly "understand" content versus merely replicating it.

Application in eDiscovery:

- The authors demonstrate LLMs' capability to revolutionize document review, analysis, and transcript review by rapidly identifying relevant information, significantly cutting down the time and resources required for these tasks.

- Through practical examples, the article showcases how GenAI can offer immediate positive effects on discovery practices, despite the architectural limits of current models.

Data Security: Addressing concerns around data privacy and the protection of attorney-client privileges, the guide reassures readers about the robust security measures and contractual safeguards in place with commercial LLM providers.

Impact and Future Outlook: The conclusion reiterates GenAI's revolutionary potential in legal workflows and its capacity to make existing processes more efficient. The

MERLIN

authors express enthusiasm for the future, envisioning a legal practice landscape redefined by the adoption of advanced AI tools.

*The above summary was prepared by GPT 4.0 Turbo (128K) based on the contents of our article.

Let's begin.

What is GenAI?

Generative AI refers to a type of artificial intelligence that can generate new content–text, images, music, or other forms of media–based on its training and the input it receives. A GenAI model is typically trained on a vast amount of text data, much of it taken from the Internet. That training, which involves reading billions and even trillions of text examples, is supplemented by thousands of hours of human interaction focused on asking the model questions and providing feedback on its answers. This process is called reinforcement learning and is critical to the model's fluency and effectiveness.

Because of the extensive training required for a GenAI model, they are often called Large Language Models ("LLMs")

What is GPT?

One of the leading forms of GenAI is called GPT, which stands for Generative Pre-trained Transformer. Although the reference is generic, OpenAI gave the name "GPT" to the LLM it created. OpenAI's GPT model is the brains behind a program called Chat GPT, which caught the world's attention in early 2023. Different versions of the GPT model are referred to by numbers, e.g. GPT 3.5, GPT 4.0, and more recently by the model's size or capability, e.g. GPT 4.0 Turbo (128k). We will have more on the parenthetical numbers in a minute.

There are hundreds, if not thousands of Generative Pre-trained Transformer models in existence today. Along with GPT itself, there is Anthropic's Claude, Google's Bard (now called Gemini), Facebook's Llama, Falcon, Mistral and many others. 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 Guide will focus on using LLMs for ediscovery

MERLIN

purposes rather than on which LLMs and which forms of delivery are better suited for your particular needs.

We will also use GPT as a more generic reference for the many different LLMs that are available, specifying OpenAI's GPT product by model number when the reference is meant to be specific.

How Do LLMs Work?

LLMs like GPT require a massive amount of computing power and run on a large collection of expensive, specialized chips called GPUs or graphical processing units. Some have suggested that OpenAI's GPT model costs hundreds of thousands of dollars a day to run. You can think of these LLMs as supercomputers, but with depth, breadth and power unlike anything that has come before.

Training

As mentioned, the LLM must be trained on a massive amount of mostly Internet text including books, articles, websites and other textual sources. This process allows it to "understand" grammar, context, and a wide variety of topics. We put the word in quotes because there is an ongoing debate about whether the LLM understands anything. Some critically 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 any event, the prediction process is intensively mathematical, requiring a huge amount of computing power at least for the larger commercial models. Efforts are underway to develop models that can run on a smaller number of servers or even on a laptop. Apple is reportedly working on models that can run, either in whole or in part, on your mobile phone.

Ultimately, the goal in training is to teach the model to predict the next word in a sentence based on the words that have come before.

Training Cutoff: The key here is that LLM training has an end point, often called a cutoff. Once training completes, the model's parameters are fixed and it can no longer learn from new data.

Whatever the model has learned up until the training cutoff becomes its permanent knowledge base. This means it won't adapt to or reflect any changes or new information in the world that occur after this point.

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. This is a key point to understand when you are talking about data security for these models. The model cannot use your prompt to broaden its knowledge base, nor can it inadvertently pass prompt information to other users.
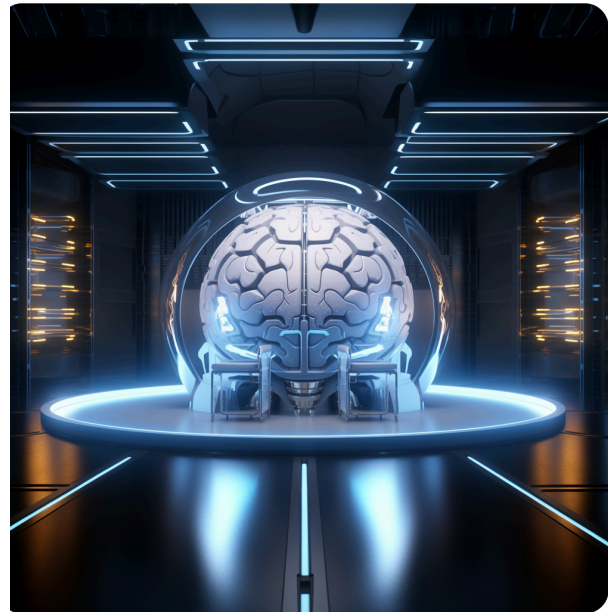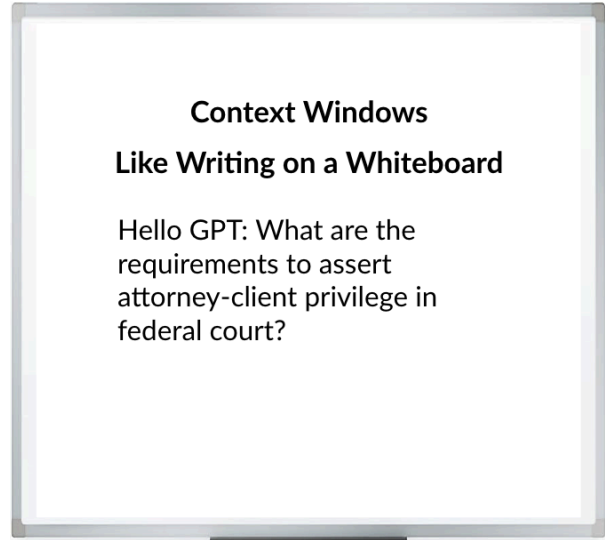
GPT is like a Brain in a Jar.

The picture above was actually created by another form of GPT, one that can create images based on your prompts rather than respond with text.

How Does GPT Communicate?

Likely most of us have interacted with ChatGPT, carrying on conversations that are eerily reminiscent of human interactions. If GPT has no memory, how can that happen?

The answer is simple but important for us to understand. GPT communicates with us through what is called a "context window." In our discussions we liken it to a white board, one that exists outside the jar but is accessible to GPT. We send text to GPT on the whiteboard. It can respond to our questions there as well. Once the answer is given, the whiteboard is erased, much like a computer's memory is erased when you turn it off.

So, when we ask GPT to write a poem, that request, also known as a "prompt," is sent to the whiteboard. GPT reads the prompt, analyzes the request and sends its answer back to the whiteboard. From there it is sent to your browser or whatever software application you are using to access GPT. And then, once the answer is given, everything is erased.

Context Windows

Like Writing on a Whiteboard

Hello GPT: What are the requirements to assert attorney-client privilege in federal court?

A Simple Context Window Whiteboard).

Carrying on a Conversation

So how does GPT carry on a conversation if it can't remember anything? Once again, the answer is simple and somewhat unexciting. A software program like ChatGPT keeps track of your conversation and sends it up 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 conversation.

In that regard it is important to understand the difference between ChatGPT and GPT. ChatGPT is a software application that OpenAI created to allow people on the Internet to communicate with GPT, the underlying LLM that analyzes and responds to your questions. The Chat part of ChatGPT saves your discussion, at least for that session, so you can continue your conversation. You may note that ChatGPT allows you to see different sessions and to revisit prior conversations even from days or weeks before. GPT can carry on the conversation as if there had been no gap in time because the Chat application resends it along with your next question or prompt.

You can thus carry on a conversation at least until the context window (whiteboard) runs out of room, at which stage ChatGPT will "forget" the early parts of the conversation.

The Context Window Size is Limited

The most important thing to know about Context Windows is that the amount of text you can place on them (prompt plus answer) 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 Guide). Thus, your conversation with GPT–including its answer–was limited to the size of the context window. When your conversation got larger than the window allowed, ChatGPT would cut out the first part of the conversation so you could continue to ask new questions. You might have noticed that GPT began forgetting aspects of your earlier conversation once you moved further in your conversation.

You can quickly 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. And likewise you couldn't and still can't ask GPT to read and analyze millions (or even thousands) of your discovery documents.

In short order, LLM capabilities increased, moving from 4 to 8 to 16 and even 32k context windows. Last summer Anthropic (founded by people from OpenAI) released a 100k version of its LLM called Claude 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. That got us excited as we watched OpenAI respond with GPT 4 Turbo (128K) and Anthropic come back again with Claude 2.2 (200K).

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 is not feasible, either technically or for cost considerations. Even if the windows can be made larger (which of course they will be), there is current concern that the models cannot actually remember everything read in large context windows, which may mean that they will overlook important details when giving their answer.

All we can say at this point is that the larger context windows open the door to use these powerful GenAI models for a variety of uses including, Ta-Dah, for ediscovery.

Is the Data We Send to GPT Secure?

We have written an entire article on this subject and given two webinars to date, 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?](#) (Law 360, 8/17/23). You can also watch [both the U.S. and London webinars here](#).

So, are we risking a waiver of attorney-client or work-product privileges by sending our data to an LLM? No, you're not taking a risk, at least not if you are using a commercial license for the service. Microsoft and the other major large language model providers 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.

Using LLMs for Ediscovery

Now that we have a good understanding of GenAI and how these LLMs work, let's tackle the heart of our subject, using LLMs to improve ediscovery workflow. In that regard, we don't plan to cover every possible use of an LLM for ediscovery, let alone for general legal purposes. Rather, by focusing on several examples we will give you an idea of how these LLMs, despite the size limitations of their Context Windows, can transform ediscovery workflow to make it more efficient and cost-effective. If your field is litigation, this will hopefully give you ideas on how you can use LLMs to improve investigation and discovery workflow. If your practice involves other aspects of the law, this might give you analogous ideas on how to better find and analyze documents to improve your workflow process.

1. Using LLMs to Review and Classify Individual Documents

A first and most obvious use of an LLM is to review and classify discovery documents. In that regard, our goal is to explain how you can use an LLM for review despite its technical limitations. We are not jumping into the argument of whether they can meet or exceed human review (other than their obvious advantage in speed). Rather, we just want to show you how they can be used as a technical matter. Whether you should or should not use an LLM to supplement human review is a matter for a different article.

The process to review and classify individual documents is rather simple. You start with a prompt which describes the purpose for the request (say, responding to a request for production) and your classification criteria. Then you include the text of the document and close with your request. Think of it as sending both your information request along with the text of the document under review to the whiteboard (context window).

The LLM reads your prompt and any information you provide about your information need and then reads the document text and your specific request. After a few seconds for analysis, the LLM responds by classifying the document, e.g. responsive or not, and, if requested, giving you a reason for its decision. Once that work is done, it moves to the next document.

For single document review, the limited size of the context window is rarely an issue, particularly with the more recent LLM versions like GPT 4.0 Turbo (128K) or Claude Instant (100K) or even Claude 2.2 (200K). These models have sufficient context window sizes that they can hold the text of your prompt, the document under review and its answer. If the document text is longer than will fit on the context window, you can decide to split up the document into sections for review or just submit the first XXX thousand tokens of the document text to the LLM.

Hey GPT: What is this about?

From: Hukill, Dorothy [mailto:Dorothy.Hukill@myfloridahouse.gov]
Sent: Thursday, March 10, 2005 4:37 PM
To: Jeb Bush
Cc: chris.flack@myflorida.com
Subject: Slot Machines- House Business Regulation Committee

Dear Governor Bush:
I am a member of the Business Regulation Committee in the House. We are in the process of discussing how much tax should be levied. The definition of "revenue" as defined by the slot proponents does not coincide with the legal definition of "revenue". They explained that revenue was what was left in the slot machines after payout. However, I believe the definition to be all money that comes into the slot machine. I brought this to the attention of the committee and I am considering referral of this for more research to our legal counsel. Considering your public position on this, I would be interested to know your thoughts.

Dorothy L. Hukill
State Representative, District 28
Cell: (386) 316-9269

Loading an Email into the Context Window.

Once the individual document review is completed, you start from scratch with a new document. Thus, individual document review is fairly easy for an LLM and the few seconds it takes the LLM to complete the process is at least 20-30 times faster than for a human reviewer.

Here is a simple example from a review of documents relating to protecting manatees in Florida. The topic was one used for the annual text retrieval conference put on by NIST to test an algorithm's ability to quickly find relevant documents.

BegControl: Bush710178
Relevance: 100

Email from Jeb Bush to Bunny Hanley with Numerous Attachments About Manatee Protection Plans in Lee County, Florida

This document is an email sent on September 23, 2004 from Jeb Bush to Bunny Hanley at the email address bunny.hanley@myflorida.com. It contains numerous attachments related to manatee protection plans and policies in Lee County, Florida.

The attachments include:

Manatee protection plan (MPP) for Lee County
Manatee protection in Lee County
Lee County MPP

MERLIN

Opposition to Lee County MPP
Protect wildlife in your state
Lee County Manatee Protection Plan
Signed resend of request
Manatee Protection
STOP THE MPP
Manatees
Please protect manatees
Manatee protection plan
Save the Manatees
Manatee Protection Plan for Lee County
MMP violations in Lee County
Saving the Manatees
Manatee Protection
Approving a manatee protection plan – HELP!
Manatees
Endangered Species Act and the Marine Mammal Protection Act
Save the manatees NOW
Lee Co. manatees
Protect Precious Manatees!!!
Manatee Protection in Lee County
Manatee Protection
SAVE THE MANATEES
Please save the manatees!
Manatee Protection
Please save the manatees
Manatees
Manatee protection?
Manatee protection plan
Save the Manatees!
The document and all its attachments are related to manatee protection plans and policies in Lee County, Florida. This is directly relevant to the topic of documents about the "Save the Manatee" program, as manatee protection plans and policies would be part of efforts to save the manatee species.

In this case we used a simple prompt asking the Claude Instant (100K) to review and summarize the document based on the topic. As you can see, the LLM provided a thorough summary of the document—and did the work in about four seconds, including the time it spent reading the document and writing out the resulting summary.

While there is a lot more to be said about the utility of using an LLM to review documents, with pros and cons, you can see how it might become an important part of what has become a 10+ billion industry.

MERLiN®

2. Using LLMs to Analyze Groups of Documents

We can also use LLMs to review and analyze groups of documents, which is a feature we offer in DiscoveryPartner. There are different ways to accomplish this task but our approach is representative of a process called "RAG" or Retrieval Augmented Generation.

Retrieval Augmented Generation

The RAG process that is most widely used requires several steps:

1. Find a set of documents for analysis. These are typically documents that are highly relevant to your topic of interest.

    a. You can use keywords and human review to identify the likely-relevant documents.

    b. If your litigation system supports it, use algorithmic search to identify likely-relevant documents.

    c. Folder the documents as you find them.

2. Next, extract the text from the documents to prepare them for submission.

    a. The LLMs we are currently using read plain, unformatted text. This is typically extracted from native files during the processing phase and loaded into the litigation system. In most cases you can send the extracted text already in the system.

    b. The more powerful LLMs can also read images, handwriting and even transcribe audio files. We are not covering that option in this Guide because we haven't tested it yet.

3. Send an appropriate number of documents (the text) to the LLM. What is the appropriate number? It depends on which LLM you are using. Remember that you have to leave room for the prompt text (to tell the LLM what to do) and the answer text. Thus, the number will vary depending on the text size of documents (in tokens) and the room required for the prompt and the analysis returned by the LLM. A long document might not even fit in the context window!

4. The LLM then reads the documents submitted and provides its answer, usually in the form of a report.

This is the process followed by Bing when you ask GPT a question. If you watch carefully, you will see that Bing first runs a search to retrieve relevant documents for analysis by GPT. Then GPT answers your question based in large part on the content of

the documents it has reviewed. This allows GPT to go beyond its training cutoff (discussed above) to analyze and comment on information that it didn't learn during training.
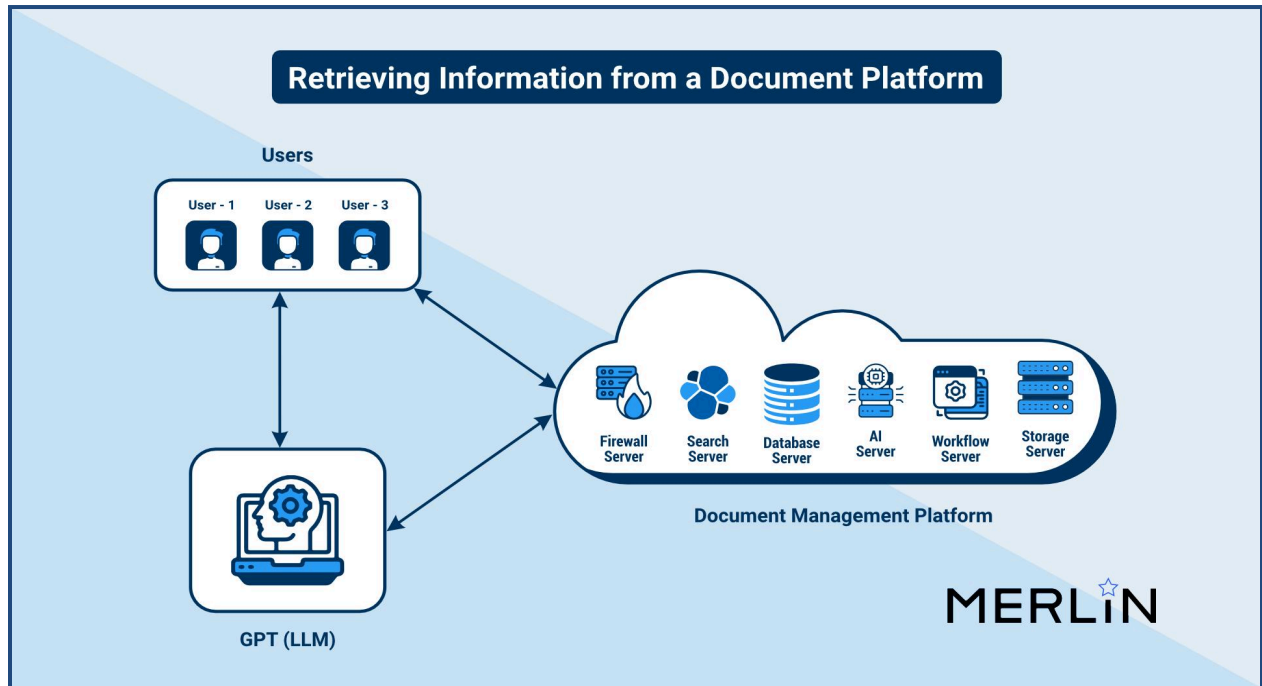
A Typical Three-Part RAG Setup

RAG can work much the same way if we want to analyze multiple discovery documents at one time. As the above illustration shows, we need an LLM-integrated litigation support system to help us find documents for submission to the LLM. When we ask the LLM a question or for a report (prompt), the LLM receives the identified documents (the text) from the litigation system for its analysis. So long as you don't send more text than the LLM's context window can handle, you will get an answer.
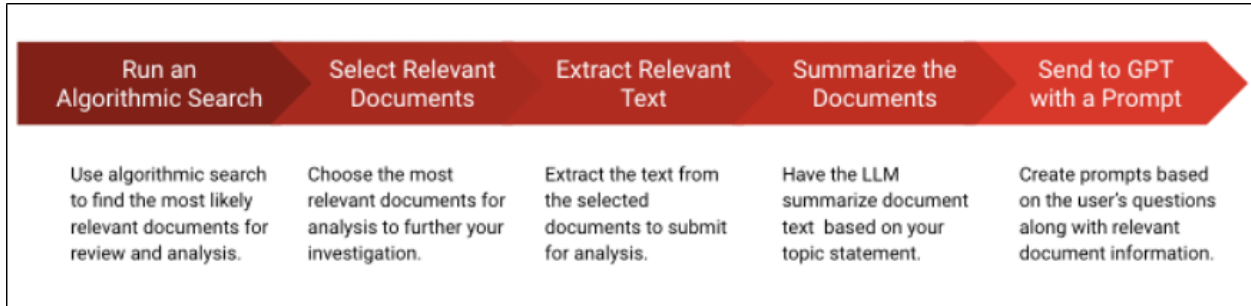
Advanced RAG

At Merlin we developed a more sophisticated RAG system for investigations and discovery. Specifically, we summarize the documents based on the prompt question (topic) before submitting them to the LLM as a group for analysis. Why do we do that? Two reasons:

1. The summaries are shorter than the original documents which means we can submit more document summaries for analysis.

2. Research suggests that the quality of the LLM's response is better when the text submitted is more narrowly focused. When whole documents are submitted, the

LLM is likely to receive a lot of extraneous content that may confuse or distract it. This is analogous to humans who can also be distracted by irrelevant content.

The modified RAG model we use looks like this:

Run an Algorithmic Search	Select Relevant Documents	Extract Relevant Text	Summarize the Documents	Send to GPT with a Prompt
Use algorithmic search to find the most likely relevant documents for review and analysis.	Choose the most relevant documents for analysis to further your investigation.	Extract the text from the selected documents to submit for analysis.	Have the LLM summarize document text based on your topic statement.	Create prompts based on the user's questions along with relevant document information.

Merlin's Advanced RAG Workflow

With the larger context windows found in GPT 4.0 Turbo (128K) or Claude (200K), we can often submit several hundred document summaries for analysis. In sharp contrast, a typical RAG system is limited to reading a much smaller number of documents.

Here is an example of an analysis on the use of eminent domain proceedings for the development of commercial properties in Florida like shopping centers.

Use of Eminent Domain for Private Development

- Pro: The Supreme Court's decision in Kelo v. City of New London expanded the definition of "public use" to include economic development, potentially allowing for increased tax revenue and economic growth (Bush1201155).

- Con: Many individuals, including Mark Lee, Jim Sebest, and Larry Walters, argued that this use of eminent domain constitutes an abuse of power and infringes on the fundamental right of property ownership (Bush1244245, Bush1207856, Bush1201230).

Legality of Eminent Domain for Private Development

- Pro: The Supreme Court's ruling in Kelo v. City of New London upheld the legality of using eminent domain for private development (Bush1201155).

- Con: Several individuals, including Larry Walters and Ben Laroche, argued that this interpretation of the law pushes the country toward a violation of the Fifth Amendment, which only allows the taking of private property for public use (Bush1201230, Bush1201155).

Morality of Eminent Domain for Private Development

MERLIN

- Pro: The Supreme Court's decision in Kelo v. City of New London could be seen as a way to promote economic development and growth (Bush1201155).

- Con: Many individuals, including Jim Sebest and Sharon White, questioned the morality of using eminent domain to seize private property for the benefit of private developers or corporations (Bush1207856, Bush1250325).

Potential for Abuse of Eminent Domain Powers

- Pro: The expanded definition of "public use" could allow for more flexibility in land use and development (Bush1201155).

- Con: Numerous individuals, including Mark Lee, Gloria McCluskey, and Dana Berliner, expressed concern about the potential for abuse of eminent domain powers, particularly in light of the broad interpretation of "blight" (Bush1244245, Bush1236435, Bush1203361).

Need for Legislative or Constitutional Reforms

- Pro: Governor Jeb Bush, Chip Mellor, and Ben Laroche advocated for a constitutional amendment or statute to protect property rights and limit the use of eminent domain for private gain (Bush1241585, Bush921671, Bush1201155).

- Con: Not explicitly stated in the document summaries.

We aren't claiming the analysis is perfect or that a human couldn't create an equal or better report. But this analysis was done in seconds for a fraction of the cost required for a human legal professional to do the work.

3. Using LLMs to Analyze Transcripts and Chat Messaging

LLMs can also be a game changer for reviewing and summarizing deposition and hearing transcripts. With an advanced RAG system, LLMs can not only summarize transcript testimony, they can answer questions across dozens of transcripts, doing so in seconds. Current transcript software can run keyword searches against transcripts but they cannot answer your questions about witness testimony.

The traditional approach for dealing with transcript testimony is to create a summary, typically in a Word 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 paper summary format looked like this:

Deposition of Lawrence Stackhouse
Jan 6th Committee

Date: Tuesday, March 22, 2022
Location: Washington, D.C.
Method: via video conferencing (Webex)

Election Fraud Beliefs and Rally Attendance

11:18-11:23	His belief in election fraud is reinforced by news and local reports in Philadelphia, where he hears that people wearing Trump shirts were not allowed to vote, which he considers a form of cheating.
12:1-12:9	Stackhouse mentions seeing reports of this in Philadelphia and hearing about it from friends who experienced it.
12:14-12:19	He confirms that he saw reports on Fox Philly and on social media about people being turned away from voting polls for wearing Trump apparel.
12:20-13:6	When asked about the Philadelphia election counting livestream, Stackhouse says he did not see reporting about that but acknowledges seeing livestream videos of people being turned away at voting polls.
13:7-13:17	Stackhouse did not participate in any activities related to alleged voter fraud following the 2020 election, nor did he attend any rallies before January 6th.
13:18-13:20	He did not make any political donations to groups challenging election issues between the 2020 election and January 6th.

Deposition Summary from Testimony Given to the January 6th Committee.

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 excerpt you see above was created by GPT 4.0 Turbo in less than an hour at a fraction of the cost for a human to do the job.

MERLiN

Taking Summaries Online

Our goal was to take summaries to a higher level, placing them online with live links to the associated testimony. Here are examples from Kayleigh McEnany's testimony to the January 6th House Investigation Committee (laid out vertically for ease of reading):

---

§13: McEnany's January 6th Morning and White House Operations

Kayleigh McEnany clarifies that President Trump's comments about Liz Cheney were political, not a physical threat. On January 6th, she arrives at the White House before Trump's speech, facing heightened security. McEnany is surprised by a CNN chyron suggesting a coup, as she expects a peaceful day. She has no communication with Trump that morning and recalls no senior staff meetings, with such meetings having become less frequent post-election. The questioner begins to address a reported Trump-Pence phone call, but McEnany's response is not included.

- McEnany clarifies that when President Trump mentioned getting rid of Liz Cheney, she interpreted it as a political move, not a physical threat, because he later spoke about primary challenges (p149:l1--p149:l6)
- The questioner shifts the conversation to the morning of January 6th and asks McEnany to recount her day in detail, referencing exhibit 39, which contains her notes from that day (p149:l7--p149:l10)
- McEnany cannot recall the exact time she arrived at work on January 6th but knows it was before Trump's speech. She remembers encountering many barriers and logistical difficulties in reaching the White House due to security measures (p149:l16--p149:l22)
- When asked about an Iran issue involving the FAA mentioned in exhibit 39, McEnany recalls seeing a news story about a strange FAA transmission related to Iran, but she does not provide specifics (p149:l23--p150:l1)

```
11        But just sort of tell us how your day started in terms of going into work.     And you

12   even in exhibit 39 have a reference there to taking longer to get in because of all the

13   Trump supporters.

14        Like roughly what time did you go in, and what did you see in terms of the crowds

15   as you went in?

16        A     To the best of my recollection, I don't remember what time I arrived.     It

17   would've been before he went to give his remarks, and I just remember there being a lot

18   of barriers and it being logistically very difficult to get to the White House that day.

19        Q     Okay.    And that was because of the crowds or because of security presence

20   or both?

21        A     To the best of my recollection, it was because of security and roads being

22   shut down.    I think we had to take a longer route in my vehicle.
```

As with the last example, this online summary was created in minutes using GPT with automatically generated links to the transcript.

MERLIN

Searching Across Transcripts

The next step was to use an Advanced RAG architecture to allow a user to search across multiple transcripts to get answers to questions about witness testimony. The system runs topic searches against the transcripts and brings back relevant sections for summarization and analysis.

Based on the sections retrieved, the LLM will provide answers to questions about testimony during the deposition or hearing.

Who was at work in the White House on the morning of January 6th?

On the morning of January 6th, Donald Trump Jr. was present at the White House. He acknowledges being at an Oval Office meeting that included himself, his family, and Stephen Miller (donald_trump_jr: p81:l9--p81:l11). The annotated copy of the President's private schedule confirms the meeting at 11:10 a.m. with Trump Jr., his brother, Kimberly Guilfoyle, Ivanka Trump, Eric Herschmann, Mark Meadows, and General Kellogg (donald_trump_jr: p81:l12--p81:l17). Trump Jr. recalls being in the Oval Office that morning after he spoke at the Ellipse but before his father did (donald_trump_jr: p81:l18--p81:l25). He also remembers returning to the White House after he spoke and then going back to the Ellipse with his father when he spoke (donald_trump_jr: p82:l2--p82:l4).

Stephen Miller believes he was with Mr. Meadows briefly that morning, but any one-on-one conversation would have been in passing (stephen_miller: p161:l19--p161:l23). Miller acknowledges an entry in the President's daily diary indicating a meeting with Donald Trump Jr., Eric F. Trump, Ivanka M. Trump, Lara Trump, Kimberly Guilfoyle, and himself at 11:11 a.m., but the time is not confirmed (stephen_miller: p163:l15--p163:l20).

Ivanka Trump states that after returning to the White House from the rally at the Ellipse, she went directly to her office on the second floor (ivanka_trump: p60l13). She recalls coming back to the office as part of the collective and proceeding with various White House-related meetings (ivanka_trump: p60:l20--p60:l22).

Keith Kellogg confirms he has walk-in privileges to the Oval Office and that the White House was not fully staffed on January 6th (keith_kellogg: p73:l25--p74:l5). Before entering the Oval Office, Kellogg encounters Don Jr., Laura Trump, Eric Trump, and Kimberly, whom he greets (keith_kellogg: p74:l6--p74:l14). After checking with the National Security Advisor's office, Kellogg proceeds to the Oval Office (keith_kellogg: p74:l23--p75:l1).

Kayleigh McEnany cannot recall the exact time she arrived at work on January 6th but

---

knows it was before Trump's speech. She remembers encountering many barriers and logistical difficulties in reaching the White House due to security measures (kayleigh_mcenany: p149:l16--p149:l22). She states that to her recollection, there were no senior staff meetings that day (kayleigh_mcenany: p150:l22--p150:l25).

Benjamin Williamson does not recall being in the Oval Office that morning before the rally on the Ellipse (benjamin_williamson: p37:l5--p37:l7). He also confirms he did not consider resigning on January 6th (benjamin_williamson: p91l4).

Chad Wolf, while not at the White House, was engaged in his capacity as Acting Secretary of Homeland Security, coordinating efforts from Qatar due to the time difference (chad_wolf: p41:l25--p42:l1).

John McEntee might have gone into the Oval Office early in the morning on January 6th if someone needed something before the President arrived, but he does not think he did that day and was definitely not there during the working day (john_mcentee: p160:l15--p160:l18).

Once again, this summary of information was created in seconds, this time based on information quickly obtained across multiple transcripts.

Conclusion

We wrote this paper to provide a Guide for smart legal professionals who want to better understand the strengths and weaknesses of Generative AI. We added several ediscovery workflow examples to provide a practical look at how GenAI can revolutionize discovery workflow–even with its architectural limits. Our goal wasn't to identify every possible way an LLM could make discovery workflow more efficient and effective but rather to offer a few examples that could immediately have a positive effect on a discovery practice.

From our perspective, the integration of GPT into the discovery workflow marks a significant advancement for the profession. From the buzz at the recent LegalWeek conference to the discussions that currently dominate legal forums, the anticipation and integration of Generative AI tools like GPT signal a paradigm shift in how legal work will be approached going forward.

In the end, this new form of AI technology offers the potential to dramatically improve the efficiency and accuracy of critical tasks such as document review, document analysis, and transcript review. By leveraging GPT's capabilities, lawyers can quickly review and analyze documents and other data, identifying relevant information with all but unbelievable speed and precision. Legal teams can thus make better use of their

MERLiN

time and resources, devoting more of their attention to helping develop trial and settlement strategy, exercising judgment and giving clients sound advice about their legal rights.

Ultimately, GenAI's promise is not just to make existing processes more efficient. Rather, we are most excited about imagining what a discovery practice can become with these new capabilities. We know that the profession is not filled with dummies but with smart, forward-thinking individuals ready to take advantage of these new tools. Together, we stand at the threshold of a new era in legal technology, one that will redefine the contours of legal work in ways we are only beginning to understand.

Key GenAI Terms Smart People Should Know

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.

Generative AI refers to 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. The name is often shortened to **GenAI.**

GPT stands for **Generative Pretrained Transformer**. It is 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 is the name given to a web-based application that allows users to talk to 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.

Large Language Model (LLM) is the name given to GenAI systems (often called models) like GPT, Claude, Bard, 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 and the volume of data they have been trained on.

Prompt: The name prompt refers to 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. The model processes this input, leveraging its training on vast datasets, or the information provided in the prompt, to generate a response that aligns with the context and content of the prompt.

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 equates to about 1,000 tokens.

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.

About Merlin Search Technologies

MERLIN

Merlin is a pioneering cloud technology company leveraging generative AI and cloud technologies to re-engineer legal investigation and discovery workflows. Our next generation platform integrates GenAI and machine learning to make the process faster, easier, and less expensive. We've also introduced Cloud Utility Pricing, an innovative software hosting model that charges by the hour instead of by the month, saving clients substantial savings on discovery costs when they turn off their sites.

With over twenty years of experience, our team has built and hosted discovery platforms for many of the largest corporations and law firms in the world. Learn more at merlin.tech.