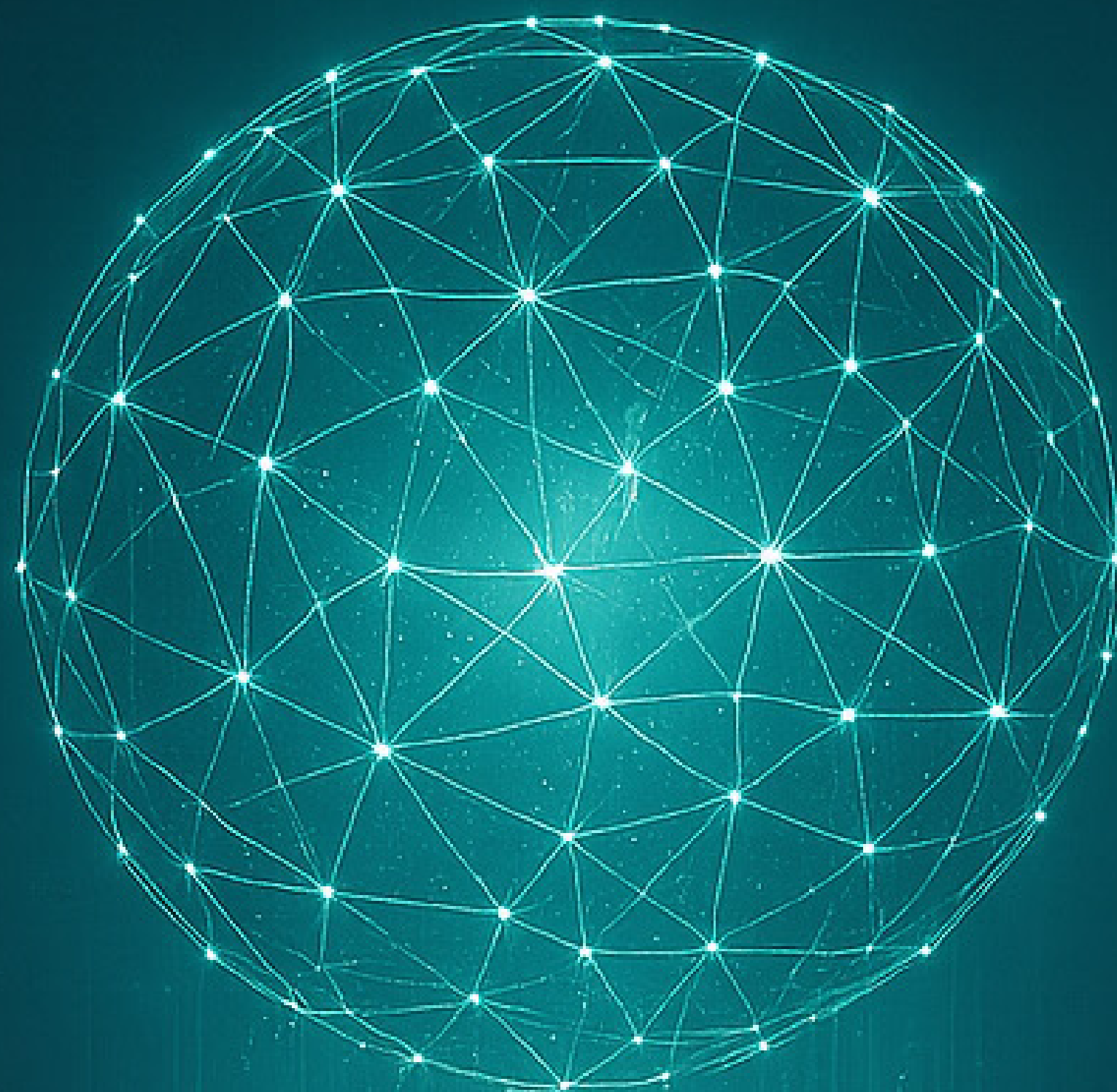


MERLIN[☆]

GENERATIVE AI



FOR SMART DISCOVERY PROFESSIONALS

4TH EDITION

JOHN TREDENNICK, J.D. WILLIAM WEBBER, PhD.



GENERATIVE AI

FOR SMART DISCOVERY PROFESSIONALS

4TH EDITION

© 2025 Merlin Search Technologies, Inc. All rights reserved.

This electronic book may be freely shared and distributed for noncommercial purposes only.
You may not use this work for commercial purposes without prior written consent.

This work is licensed under a Creative Commons Attribution-NonCommercial- ShareAlike 4.0 International License. For details,
visit <https://creativecommons.org/licenses/by-nc-sa/4.0/>

FOREWORD TO GENERATIVE AI 4TH EDITION

By William Hamilton

Introduction to the Fourth Edition

For more than two decades, the e-discovery field has been plagued by a quiet yet persistent divide—an artificial and ultimately unhelpful compartmentalization between e-discovery, litigation, and trial practice. For twenty years, legal technology and ESI promised to unite these silos and provide us with understandable information grounded in the concrete immediacy of electronically stored information. That promise was never fully realized. The tools were powerful, but partial; the analytics were helpful, but narrow; the workflows sophisticated, but never truly integrated into the larger craft of discovery, analysis, and advocacy.

John Tredennick’s and William Webber’s *Generative AI for Smart Discovery Professionals* (4th ed.) arrives at the exact moment when the field is ready—perhaps for the first time—to see e-discovery, litigation, and trial practice not as parallel tracks but as a unified whole. This book is comprehensive, clear, and deeply practical. It cuts through mythology, marketing noise, and unfounded anxiety. Most importantly, it demonstrates that modern AI does not merely accelerate e-discovery; it restores the lawyer’s capacity to understand, reason, argue, and tell the story of a case from beginning to end.

This book shows us the open door to an era in which AI makes the litigator whole again.

Part One: Explaining What Was Never Fully Understood

Foundations of Generative AI for Discovery Professionals

Part One is indispensable. Before this book, most discovery professionals had some familiarity with large language models (LLMs), but very few understood what they actually are. The authors remedy this deficiency and succeed in explaining, in plain but technically accurate prose, how LLMs work, why context windows matter, why hallucinations happen, and why security concerns must be understood rather than feared.

Consider Chapter 1’s explanation that LLMs are “brains in jars” without memory, whose only interface with the world is a temporary “whiteboard” onto which users write their instructions. This metaphor, simple yet powerful, clarifies the central mechanism of AI: the context window. Lawyers finally gain an accessible understanding of how these systems process information and why they cannot “remember” anything unless that information is placed before them each time.

Chapter 2’s discussion of the shift from 4k-token to 128k-token and larger context windows highlights the book’s focus on clear teaching. Instead of following hype, the chapter explains real limits: why bigger windows are important, when they aren’t, and the well-documented “lost in the middle” issue that restricts how effectively models utilize large contexts.

Chapter 3’s treatment of hallucinations—complete with examples from the *Mata v. Avianca* sanctions and other recent cases—is especially valuable. The book emphasizes that hallucinations are not mysterious failures but predictable results of how LLMs generate text. Lawyers and judges will find here the clearest roadmap yet for preventing AI errors and establishing

verification workflows that align with legal ethics and professional responsibility standards.

Part One concludes with clear discussions of confidentiality, privacy, and ethical obligations. The tone is serious but motivating: AI is powerful, but using it responsibly depends on competence, verification, and professional judgment. In other words, it requires lawyering across discovery, analysis, and advocacy alike.

Part Two: Agentic AI, Reasoning Models, and the Next Evolutionary Step

From Prediction to Purpose: How Agentic AI Plans, Reasons, and Executes Legal Work

If Part One explains LLMs as they are, Part Two explains what they are becoming.

This section is one of the book's most important contributions. Here, Tredennick and Webber describe agentic AI—systems that can break tasks into subtasks, execute multi-step plans, revise their own work, and collaborate with other models. This is not science fiction; it is happening now. Agentic systems already “plan multi-step investigations, analyze patterns across thousands of transcripts, and collaborate with other AI systems to tackle complex discovery challenges.

Chapter 6 explains how reasoning models perform multi-step analysis. Rather than offering abstractions, the authors show how modern LLMs decompose legal questions, test hypotheses against source material, and revise their conclusions through iterative passes. The book demystifies this capability and makes clear its value: litigation is reasoning under pressure, and AI now assists not merely with text but with reasoning.

Chapter 7 walks through agentic workflows in discovery—an area where practitioners badly need guidance. The book is refreshingly candid: automation without human oversight is dangerous, yet failing to leverage automation is professionally irresponsible. The authors show how to balance both.

Chapter 8 concludes Part II by examining the ethical implications of agentic systems. Here, the authors are at their best, insisting that transparency, auditability, and accountability are not optional features but foundational principles. This chapter should be required reading for any lawyer who intends to deploy AI in matters of consequence.

Part Three: The Lawyer Reunited—Practical Use of AI in Review, Argument, and Trial

(Practical Applications in Discovery and Investigations)

Document Review and Deep Analysis

Part Three delivers what practitioners need most: concrete demonstrations of AI at work across the full arc of discovery, analysis, and advocacy.

Chapter 11 demonstrates how AI now interprets documents with context, nuance, and narrative flow—something keyword search and early predictive coding never accomplished. The examples are impressive. The AI finds themes across documents, connects them to timelines, and uncovers patterns that no reviewer could identify alone. This isn't replacing human reviewers; it's finally expanding their analytical capabilities as they were always meant to have.

Transcripts, Medical Records, and Evidence Synthesis

Chapter 12 walks through transcript analysis, highlighting how AI can extract witness concessions, contradictions, and factual themes across hundreds of pages with accuracy and source citations. This is exactly the kind of practical, grounded application litigators have lacked until now.

Building Timelines and Chronologies

Chapter 13 shows how AI constructs case timelines by linking documents, emails, and testimony into coherent, cited narratives—a task that has always strained human memory and attention. The examples bring the workflow to life, illustrating how an AI can meaningfully transform early case assessment.

Drafting Arguments and Cross-Examination

Chapter 14 is perhaps the most consequential: it shows how AI helps craft arguments, develop cross-examination outlines, and write persuasive narratives—all based on the underlying documents and transcripts. The model does not replace advocacy; it strengthens it. It gives litigators a unified view of the case, allowing argument, discovery, and trial preparation to be aligned rather than siloed.

In these examples, the promise of AI today finally comes into view. AI does not fragment the lawyer's role; it reintegrates it. The lawyer becomes again what trial practice demands: a master of the facts, interpreter of evidence, and storyteller of the case.

Closing an Epoch and Opening a New One

This book marks the end of an era. For twenty years, e-discovery lived with a fracture—between search and story, between analytics and advocacy. The tools were good, but they were never whole. AI changes that.

The book shows that we are standing at the close of a fractured epoch. AI has finally given us the tools to be a unified e-discovery, litigation, and trial lawyer: a complete practice where discovery, investigation, reasoning, and trial strategy converge again.

John Tredennick and William Webber have given the profession not only an account of where AI is, but a vision of where we are going and why the long-deferred unity of discovery, analysis, and advocacy is finally becoming a lived reality.

William F. Hamilton
Master Legal Skills Professor
University of Florida Levin College of Law
November 2025

A NOTE FROM THE AUTHORS



John Tredennick, J.D.



William Webber, PhD.

Introduction to the Fourth Edition

When ChatGPT launched in November 2022, it signaled the arrival of a new technological era. Now, several years into the Generative AI revolution, we can assess its impact with clearer eyes. What once seemed like science fiction has become standard practice in many legal workflows. Discovery professionals who dismissed AI as hype now find themselves working alongside these systems daily. The question is no longer whether AI will transform legal work, but how to harness its capabilities effectively while maintaining professional standards and ethical obligations.

This transformation extends far beyond simple document search or basic summarization. Today's AI systems can plan multi-step investigations, analyze patterns across thousands of transcripts, extract structured data from unstructured documents, and even collaborate with other AI systems to tackle complex discovery challenges. These “agentic” systems represent a fundamental shift in how we approach information analysis. Rather than simply responding to our prompts, they can pursue goals, iterate on their approaches, and refine their work through multiple passes.

For discovery professionals, this evolution matters immensely. The volume and complexity of data in modern litigation continues to grow exponentially. Traditional approaches struggle to keep pace. A single matter might involve millions of documents, hundreds of custodians, dozens of languages, and terabytes of data spanning email, chat, social media, and proprietary systems. Manual review remains prohibitively expensive. Keyword search misses critical context. Even traditional predictive coding has limitations when dealing with nuanced questions that require synthesis across documents.

Generative AI offers something different. These systems can read documents much as humans do, understanding context and nuance. They can follow complex instructions, synthesize information from multiple sources, and present findings in clear narrative form. More importantly, when properly designed and deployed, they can do this at scale while maintaining audit trails and providing citations to source material.



But this power comes with responsibility. Discovery professionals must understand not just what these systems can do, but how they work, where they fail, and what safeguards are necessary. We must grapple with ethical questions about confidentiality, accuracy, bias, and professional competence. We must develop new workflows that combine AI efficiency with human judgment. And we must stay current as technology continues its rapid evolution.

Our Perspective

We write from direct experience building and deploying AI systems for discovery and investigations. We have seen these systems succeed brilliantly and fail unexpectedly. We have worked with clients who embraced AI early and others who approached it cautiously. We have learned from both successes and mistakes.

This practical experience shapes our perspective. We are neither AI evangelists who see it as a panacea nor skeptics who dismiss its potential. We recognize that AI represents genuine innovation in legal technology while acknowledging its current limitations. We believe discovery professionals can and should leverage these capabilities, but only with proper understanding, appropriate safeguards, and clear-eyed assessment of risks.

Our goal is to equip you with the knowledge needed to make informed decisions about AI in your own practice. We want you to understand the technology well enough to evaluate vendor claims critically, implement AI systems effectively, and use them ethically. We want you to see opportunities others might miss while avoiding pitfalls others might stumble into.

Most importantly, we want you to approach AI as a tool that enhances your professional capabilities rather than replaces them. These systems amplify human expertise—they don't substitute for it. Your judgment, experience, and ethical grounding remain essential. AI changes how you work, not whether your work matters.

Welcome to the evolving world of AI-powered discovery. Let's explore it together with the rigor, skepticism, and intellectual curiosity the subject deserves.

John Tredennick and William Webber

November 1, 2025

TABLE OF CONTENTS

PART
ONE

Foundations of Generative AI for Discovery Professionals

Introduces the essential concepts and principles for using AI responsibly in discovery.

CHAPTER
01

Understanding Large Language Models

Explains how LLMs learn, reason, and generate language for legal work.

CHAPTER
02

Context Windows and Conversation Management

Describes how AI maintains context without long-term memory.

CHAPTER
03

Hallucinations and Accuracy

Examines why AI produces errors and how to ensure factual reliability.

CHAPTER
04

Data Security, Privacy, and Confidentiality

Outlines safeguards to protect sensitive client data in AI systems.

CHAPTER
05

Ethical and Professional Responsibilities

Defines how legal ethics apply to AI-assisted discovery practices.

PART
TWO

Agentic AI and Reasoning Models

Explores intelligent AI systems capable of reasoning, planning, and acting autonomously.

CHAPTER
06

Understanding Agentic AI Systems

Introduces AI agents that plan and execute complex legal workflows.

CHAPTER
07

Reasoning Models and Deep Analysis

Explains how advanced AI models perform multi-step logical analysis.

CHAPTER
08

Agentic Workflows in Discovery

Shows how to balance AI automation with human oversight and control.

CHAPTER
09

Ethical Considerations for Agentic Systems

Addresses accountability, bias, and transparency in autonomous AI.

TABLE OF CONTENTS

PART
THREE

Practical Applications in Discovery and Investigations

Demonstrates real-world use cases of AI improving efficiency and insight in discovery.

CHAPTER
10

Using GenAI for Investigations and Early Case Assessment

Accelerates case analysis through intelligent evidence review.

CHAPTER
11

Using GenAI for Document Review and Analysis

Automates large-scale document review with contextual understanding.

CHAPTER
12

Using GenAI for Transcripts and Medical Record Analysis

Extracts critical facts and insights from complex records.

CHAPTER
13

Using GenAI to Build Timelines and Chronologies

Creates accurate case timelines by connecting key data points.

CHAPTER
14

Using GenAI to Write Powerful and Persuasive Arguments

Assists in crafting clear, evidence-based legal narratives.

CHAPTER
15

Building AI-Augmented Discovery Practice

Guides integration of AI tools into modern discovery operations.

APPENDIX
ONE

GenAI Terms Smart Legal Professionals Should Know

Defines essential AI terminology for legal professionals.

APPENDIX
TWO

Prompt Engineering Guide for Discovery Tasks

Explores how AI will continue shaping the future of discovery.

APPENDIX
THREE

About the Authors

Highlights the authors' backgrounds and AI experience at Merlin Search Technologies.



PART ONE

FOUNDATIONS OF GENERATIVE AI FOR DISCOVERY PROFESSIONALS

PART ONE

Foundations of Generative AI for Discovery Professionals

Before you can deploy Generative AI effectively in discovery and investigations, you need to understand what these systems are, how they work, and what constraints govern their use. Part One provides this essential foundation.

Many legal professionals approach AI with a mix of curiosity and concern. The technology seems powerful but opaque. Headlines tout revolutionary capabilities while also warning about hallucinations and security risks. Marketing materials promise dramatic efficiency gains, but ethical guidance counsels caution. This combination often leaves discovery professionals uncertain about how to proceed.

Part One cuts through this confusion by providing clear, accurate explanations of the technology underlying Generative AI. We explain not just what Large Language Models can do, but how they do it and why that matters for your work. We address the most serious concerns about accuracy and security directly, examining both genuine risks and overblown fears. And we provide the ethical framework needed to use these tools responsibly within the bounds of professional obligations.

What You'll Learn

Part One establishes the technical foundation you need, covering five essential topics: how Large Language Models actually work, what context windows are and their implications for discovery workflows, why hallucinations occur and proven mitigation strategies, whether client information remains secure in AI systems, and what ethical obligations govern AI use in legal practice. You'll understand not just what these systems can do, but their fundamental limitations, security characteristics, and the professional responsibilities that constrain their deployment.

Why This Foundation Matters

You might be tempted to skip directly to Part Three's practical applications and examples. We understand the impulse. Discovery professionals face immediate challenges requiring practical solutions. Reading about transformer architectures and training cutoffs might seem academic when you need to review a million documents by next month.

But experience teaching legal professionals about AI has taught us that understanding fundamentals pays dividends. Those who grasp how LLMs work make better decisions about when and how to deploy them. They design more effective workflows. They catch errors others miss. They explain their methodologies more convincingly to clients, courts, and opposing counsel. And they adapt more successfully as the technology evolves.

Consider hallucinations. If you understand why they occur—that LLMs predict probable next

words rather than retrieve verified facts—you'll naturally implement appropriate verification. If you view hallucinations as mysterious malfunctions, you might either avoid AI entirely or trust it naively. Neither serves your clients well.

Similarly, understanding context windows explains why some discovery tasks work beautifully with AI while others require different approaches. Knowing that LLMs have no memory clarifies both security characteristics and practical limitations. Grasping the ethical framework allows you to adopt AI confidently rather than hesitantly, knowing you're meeting professional obligations.

How to Approach Part One

We've written Part One to be accessible regardless of technical background. We explain concepts clearly without assuming prior AI knowledge. But we don't oversimplify. Discovery professionals deserve accurate understanding, not vague analogies that break down under scrutiny.

If you find certain sections challenging, persist through them. The investment pays off. If particular topics interest you deeply, use them as starting points for further exploration. We provide frameworks for understanding AI, not exhaustive technical documentation.

Most importantly, read Part One with your practice in mind. As you learn about context windows, consider your typical document volumes. When we discuss hallucinations, think about your verification workflows. While exploring ethics, reflect on your client relationships and professional obligations. The goal isn't abstract knowledge but practical understanding that improves your work.

Looking Ahead

Part One establishes the foundation for everything that follows. Part Two will build on this foundation, exploring agentic AI systems that can plan, execute, and refine complex multi-step workflows. Part Three will demonstrate these capabilities through detailed discovery applications, showing how the principles from Part One translate into practical tools that transform legal work.

But those advanced capabilities and practical applications rest entirely on the fundamentals we cover here. Understanding what LLMs are, how they process information, where they fail, how to protect confidential data, and what ethics requires provides the knowledge base for responsible, effective AI adoption.

Let's begin with the most basic question: What exactly is a Large Language Model, and how does it work?



01

Understanding Large Language Models

CHAPTER ONE

Understanding Large Language Models

The technology behind Generative AI represents one of the most significant developments in computing history. For discovery professionals, understanding how these systems work is not merely academic curiosity. It provides the foundation for using them effectively, recognizing their limitations, and making informed decisions about when and how to deploy them in legal work.

This chapter introduces the fundamental concepts underlying Large Language Models. We explain what these systems are, how they're built, and why they matter for legal practice. Our goal is not to make you a machine learning engineer, but to give you sufficient understanding to use these tools confidently and critically.

What is Generative AI and How Does it Work?

Generative AI is artificial intelligence capable of creating new content—text, images, videos, audio, code, and other data forms. Unlike earlier AI systems that could only classify or analyze existing information, generative systems produce original content with similar characteristics to their training data.

This capability matters tremendously for discovery work. Traditional software can search documents for keywords, count their occurrences, and organize results. But it cannot read a document and explain what it means in context. It cannot synthesize information across hundreds of sources to answer a complex question. It cannot draft a coherent summary that captures the essence of testimony spanning multiple depositions.



Generative AI can do all of these things. The technology achieves this through advanced neural networks called Large Language Models, typically trained on billions of text examples. Through this training, these models learn to recognize and replicate patterns and structures in language, enabling them to create content that closely resembles human-generated work.

The key insight is that language follows patterns. Words don't appear randomly in sequence—they follow grammatical rules, semantic relationships, and contextual conventions. "The judge issued a" will likely be followed by words like "ruling," "decision," or "order," not "banana" or "combustion." Large Language Models learn these patterns at a scale that allows them to generate remarkably coherent and contextually appropriate text.



At their core, LLMs perform one deceptively simple task: predicting what word should come next in a sequence based on the words that came before. This prediction process, when repeated across trillions of examples during training, produces systems capable of sophisticated language understanding and generation.

What are LLMs and Why They Matter?

Because of the extensive training and expansive computational resources required, these Generative AI systems are called Large Language Models (LLMs). The term captures two essential characteristics: they are large both in terms of their neural network architecture and the volume of data they process, and they specifically focus on language understanding and generation.

GPT (Generative Pre-trained Transformer) is an LLM, but many others now exist. The current landscape includes Anthropic's Claude, Google's Gemini, Meta's Llama, Mistral, and hundreds of other models. Some are proprietary and available only from their publishers through cloud services. Others are available through open-source licenses, allowing organizations to host them independently.

For discovery professionals, several characteristics distinguish LLMs from earlier legal technology:

Contextual understanding: LLMs don't just match keywords—they understand context, nuance, and relationship between concepts. A search for “apple” in traditional systems returns every document containing that word, regardless of whether it discusses fruit, technology companies, or the “apple of someone's eye.” An LLM understands which meaning applies based on surrounding context.

Natural language interaction: You can communicate with LLMs using plain English questions and instructions rather than Boolean search syntax or specialized query languages. This dramatically lowers the technical barrier for sophisticated document analysis.

Synthesis across sources: LLMs can read multiple documents and synthesize their content into coherent narratives, identifying themes, contradictions, and relationships that might take human reviewers days to assemble.

Adaptability: The same LLM can perform document summarization, extract structured data, answer questions, draft correspondence, or analyze contracts. This versatility makes LLMs fundamentally different from purpose-built legal software that handles only specific tasks.

Scalability: Once deployed, LLMs can analyze thousands of documents as easily as one, maintaining consistent quality across large document populations.

These capabilities make LLMs particularly valuable for discovery work, where professionals routinely confront massive document volumes, tight deadlines, and complex questions requiring synthesis across sources.



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

GPT and the Transformer Architecture

The term “GPT” stands for Generative Pre-trained Transformer. OpenAI coined this name for its groundbreaking language model, but the acronym has become somewhat generic, describing the architectural approach that most modern LLMs employ.

Let's unpack what each component means:

Generative: The model creates original text rather than simply selecting from pre-existing options. When you ask a question, the system generates its response word by word, choosing each successive word based on probability distributions learned during training.

Pre-trained: Before the model ever interacts with end users, it undergoes extensive training on vast text collections. This pre-training phase is where the model learns language patterns, factual information, and reasoning capabilities. The “pre” distinguishes this foundational training from later fine-tuning.

Transformer: This refers to the neural network architecture underlying the model. Transformers, introduced in a 2017 research paper titled “Attention Is All You Need,” revolutionized natural language processing through a mechanism called “attention.” This allows the model to weigh the importance of different words in context, understanding that in the phrase “the bank of the river,” the word “bank” relates more closely to “river” than to financial institutions.

Different versions of these models are identified by numbers and names. OpenAI's progression includes GPT-3, GPT-3.5, GPT-4, GPT-4 Turbo, and GPT-4o. The numbers indicate major architectural improvements and expanded capabilities. Letters like "o" (in GPT-4o, standing for "omni") often signify specific enhancements—in this case, multimodal capabilities handling text, images, and audio.

Competing models follow similar naming conventions. Anthropic's Claude progressed from Claude 1 through Claude 2 to Claude 3 (with variants Opus, Sonnet, and Haiku indicating different capability-cost trade-offs), and now Claude Sonnet 4 and 4.5. Google's Gemini, Meta's Llama, and others maintain their own versioning schemes.

For practical purposes, you need not master these technical distinctions. What matters is un-

derstanding that different models have different strengths, and newer versions generally outperform their predecessors. When evaluating AI tools for discovery work, pay attention to which underlying models they use and how recently those models were released.

Training Process: From Data to Intelligence

Understanding how LLMs are trained helps explain both their remarkable capabilities and their notable limitations. The training process typically involves two major stages: pre-training and fine-tuning.

Pre-training: Learning from Massive Text Collections

The first stage involves exposing the model to enormous amounts of text—typically hundreds of billions of words drawn from books, articles, websites, code repositories, and other sources. During this unsupervised learning phase, the model doesn't receive explicit instruction about what different texts mean. Instead, it learns by repeatedly predicting the next word in sequences and adjusting its internal parameters when predictions prove incorrect.

Imagine showing someone millions of legal documents and asking them to guess the next word in each sentence, providing feedback on each guess. Over time, that person would develop intuitions about legal language patterns: what terms typically follow “motion to,” how contract provisions are structured, which arguments courts find persuasive. LLMs learn similarly, but at a scale impossible for humans—processing trillions of word predictions during training.

This process is extraordinarily expensive. Training GPT-4 reportedly cost around \$100 million, including cloud computing expenses for the specialized hardware clusters required. The training consumed months of computing time on systems employing tens of thousands of powerful GPUs (graphics processing units) originally designed for rendering video but repurposed for AI calculations.

The result is a model that has internalized vast amounts of information and learned sophisticated language patterns. It “knows” (in a statistical sense) facts about history, science, law, culture, and countless other domains. It can complete sentences naturally, answer questions, and generate coherent text on diverse topics.

Fine-tuning: Refining Through Human Feedback

Pre-training alone produces models that understand language but don't always behave helpfully. They might generate toxic content, refuse straightforward requests, or provide responses that are technically accurate but practically useless. Fine-tuning addresses these issues.

During fine-tuning, humans interact extensively with the model, asking questions and providing feedback on responses. This process, often called Reinforcement Learning from Human Feedback (RLHF), teaches the model to generate responses that are helpful, harmless, and honest. Thousands of hours of human interaction guide the model toward more useful behaviors.

For legal applications, fine-tuning proves particularly important. A model trained only on general internet text might not understand professional conventions, confidentiality concerns, or

the precision required in legal contexts. Fine-tuning helps models learn these domain-specific requirements.

Some vendors create specialized legal LLMs through additional fine-tuning on legal texts. Others use general-purpose models but carefully craft prompts to invoke appropriate behavior for legal tasks. Both approaches can produce effective tools, though they involve different trade-offs in cost, customization, and maintenance.

Training Cutoff: A Critical Limitation

One crucial aspect of LLM training is the cutoff date. During training, the model learns from data available up to a specific point in time. Once training completes, the model's parameters are fixed. It cannot learn from new information or update its knowledge base.

For example, if an LLM's training data cutoff is January 2025, it will have no knowledge of events, developments, or information emerging after that date. This creates several implications for legal work:

Current events: The model cannot answer questions about recent news, new regulations, or fresh case law without additional mechanisms to access current information.

Document analysis: The model has never seen your client's documents, internal communications, or case-specific materials, regardless of when they were created. Even documents from 2020 are unknown to the model if they weren't part of its training data.

Legal research: For recent cases or statutory changes, the model lacks native knowledge and may confidently provide outdated information.

This limitation is why most practical legal AI applications employ Retrieval Augmented Generation (RAG), which we explore in Chapter 3. RAG systems provide the model with current information from external sources, overcoming the training cutoff limitation.

Understanding training cutoff also helps explain another characteristic: LLMs have no memory of conversations. The trained model is static, like a "brain in a jar" that cannot form new memories or learn from interactions. When you submit a question, the model processes it and generates a response, but it retains nothing once that response is complete. This has important security implications we address in Chapter 4.



LLMs are like a brain in a jar. They can't remember or learn from prompts.

LLMs as Modern Supercomputers

The computational requirements for running LLMs rival those of traditional supercomputers. This matters for understanding both the capabilities and the economics of these systems.

LLMs require specialized hardware called GPUs (Graphics Processing Units). Originally designed for rendering video game graphics, these chips excel at the parallel mathematical operations required for neural network computations. A single high-end GPU suitable for LLM work costs over \$100,000. The largest models require clusters of tens of thousands of these chips operating in concert.

Microsoft reportedly deployed 25,000 NVIDIA A100 GPUs for its GPT-4 implementation. The hardware alone represents hundreds of millions of dollars in capital investment. Operating costs add substantially to this figure—running an advanced LLM like GPT-4o reportedly costs over a million dollars daily in electricity and infrastructure expenses.

Why such massive requirements? The prediction process that makes LLMs valuable involves extraordinarily complex mathematical operations. Each word prediction might involve trillions of calculations. Generating a single page of text can require more computational work than running traditional database queries across millions of records.

This computational intensity has several practical implications:

Cloud delivery: Most legal professionals will access LLMs through cloud services rather than running them locally. The infrastructure costs make on-premise deployment impractical for all but the largest organizations.

Latency considerations: Even with massive computing power, generating lengthy responses takes time. When planning workflows, account for the seconds or minutes required for complex analyses.

Cost structures: AI vendors typically charge based on the amount of text processed (measured in tokens). Understanding these costs matters when designing discovery workflows that might analyze millions of documents.

Environmental impact: The energy consumption of large AI systems raises sustainability questions. Some organizations factor this into technology decisions.

Efforts continue to develop more efficient models that can run on smaller infrastructure. Apple reportedly works on models that can operate partially on mobile phones. Some open-source models can run on single high-end servers. But building and training the most capable models still requires supercomputer-scale resources.

For practical purposes, think of these LLMs as extraordinarily powerful computational resources accessible through simple interfaces. You interact with them through text, but behind that interface operates computing infrastructure of unprecedented sophistication and scale.

The Current Model Landscape

As of this writing, several major LLM families dominate the legal technology landscape. Understanding their characteristics helps you evaluate AI tools and platforms.

OpenAI GPT Series: OpenAI's models, particularly GPT-4o and its variants, remain widely used in legal applications. These models excel at complex reasoning and instruction-following. They

power numerous legal AI tools through OpenAI's API. GPT-4o offers strong performance across diverse tasks with reasonable cost-effectiveness. The "o" designation indicates omni-modal capability—handling text, images, and eventually audio in integrated fashion.

Anthropic Claude Series: Claude Sonnet 4 and 4.5 have earned strong reputations in legal circles for careful, nuanced responses. Anthropic emphasizes safety and reliability, appealing to legal professionals concerned about accuracy. Claude models generally produce fewer hallucinations than some alternatives, though no model is immune to this issue. Claude Sonnet 4.5 in particular has gained adoption for complex legal analysis tasks. Anthropic also offers extended context windows, allowing analysis of very long documents in single sessions.

Google Gemini: Google's entry into the LLM space brings substantial resources and integration with Google's ecosystem. Gemini models offer competitive performance and particularly strong multimodal capabilities. For organizations already using Google Workspace, Gemini integration provides convenient access. Gemini also offers extensive context windows approaching or exceeding one million tokens in some configurations.

Meta Llama Series: Meta releases its Llama models under open-source licenses, making them available for self-hosting and customization. This appeals to organizations with data sovereignty concerns or specific fine-tuning needs. While requiring more technical expertise to deploy, Llama models offer flexibility unavailable with purely cloud-based alternatives. Several legal tech vendors build on Llama foundations, adding legal-specific fine-tuning.

Specialized Legal Models: Some vendors have developed models fine-tuned specifically for legal work. These typically start with base models like those above but undergo additional training on legal texts and tasks. Examples include Harvey's legal-specific models and various court-specific tools. Whether specialized models outperform general-purpose LLMs with good prompting remains debated.

Model Characteristics to Consider:

When evaluating LLMs for discovery applications, several factors matter:

Reasoning capability: How well does the model handle complex logical tasks, multi-step problems, and nuanced interpretation? This varies significantly across models and versions.

Accuracy and reliability: What is the hallucination rate? How often does the model confidently assert incorrect information? This proves particularly critical for legal work.

Context window size: How much text can the model process simultaneously? Larger windows enable analyzing longer documents or more sources together, but don't assume bigger is always better—there are trade-offs we explore in Chapter 2.

Speed and cost: How quickly does the model generate responses, and what does it cost per token processed? These factors affect practical scalability for large discovery projects.

Multimodal capabilities: Can the model handle images, charts, and other non-text content? This matters when analyzing document productions containing diverse file types.

API reliability and support: For production legal use, consistent availability and responsive support matter. Downtime during critical deadlines is unacceptable.

Data handling and security: Does the vendor commit to zero-data retention? Are there adequate security certifications? We address these questions thoroughly in Chapter 4.

The Rapid Pace of Change:

The model landscape evolves faster than book publishing allows. New models emerge monthly. Capabilities improve continuously. Costs decline as efficiency increases. By the time you read this, newer models will likely have superseded those we describe.

Rather than memorizing current model names and specifications, focus on understanding the evaluation framework. When new models arrive, assess them using the criteria above. Ask vendors specific questions about reasoning capability, accuracy, security, and costs. Test models on representative tasks from your actual workflows.

The fundamental concepts underlying LLMs remain relatively stable even as specific implementations advance. Understanding how these systems work, what they can and cannot do, and how to evaluate them provides durable knowledge that survives the next generation of models.

Looking Ahead

This chapter established the foundation: LLMs are sophisticated statistical models trained on vast text collections to predict and generate language. They achieve remarkable capabilities through massive computational resources and careful training processes. Multiple competing models exist, each with distinct characteristics and trade-offs.

With this foundation, we can now explore how these systems actually interact with users and handle extended conversations. Chapter 2 examines context windows—the mechanism through which LLMs receive information and maintain apparent memory across multi-turn interactions. Understanding context windows proves essential for designing effective discovery workflows and recognizing the practical limits of what LLMs can accomplish in single sessions.

The technology we've described in this chapter enables the practical applications we explore later. But between understanding how LLMs work and successfully deploying them in discovery lies important knowledge about their operational characteristics, limitations, and risks. The next chapters build that knowledge systematically, preparing you to use these powerful tools effectively and responsibly.



02

Context Windows and Conversation Management

CHAPTER TWO

Context Windows and Conversation Management

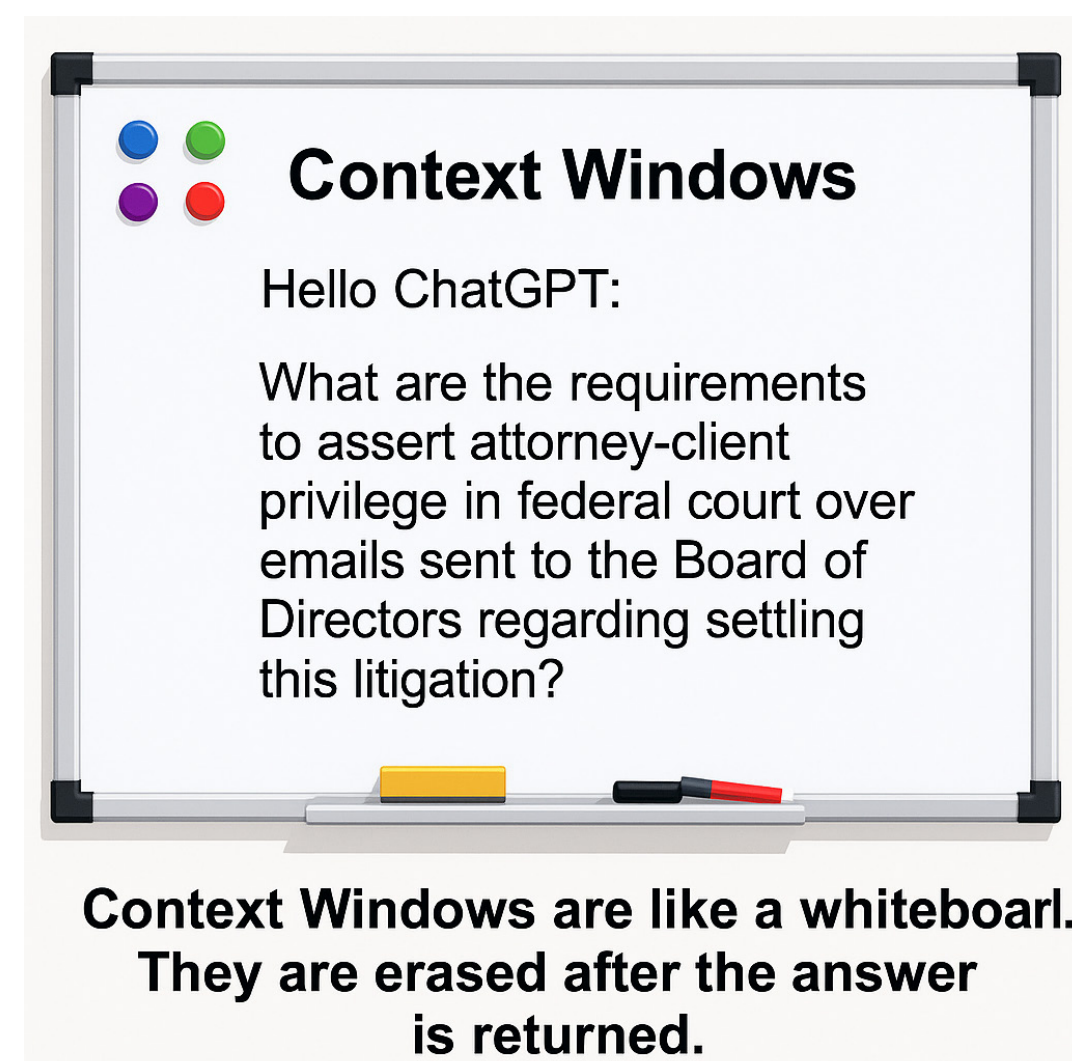
In Chapter 1, we described Large Language Models as having no memory—unable to learn from interactions or retain information after processing. Yet when you use ChatGPT or similar interfaces, you can carry on extended conversations where the AI seems to remember everything you discussed. These observations appear contradictory. How can a system with no memory maintain coherent multi-turn dialogues?

The answer lies in understanding context windows and the relationship between LLMs and the interfaces through which we access them. This chapter explores these mechanisms, explaining how AI systems manage conversations, what limitations these mechanisms impose, and what implications they hold for discovery work.

The Context Window Metaphor: The Whiteboard Outside the Brain

Recall our description of a trained LLM as a “brain in a jar”—a powerful analytical engine that processes information but retains nothing. This metaphor captures an essential truth about how LLMs operate, but it requires extension to explain how conversations work.

Imagine that the brain in a jar has access to a whiteboard positioned outside it. The brain can read what appears on the whiteboard and write responses back, but the whiteboard itself sits in external memory, separate from the brain’s internal structure. Once the brain finishes processing information and writing its response, that response is erased from its awareness. The brain returns to its baseline state, ready for the next request, with no memory of what just happened.



This whiteboard is the context window. It serves as the communication interface between the external world and the LLM’s processing core. Everything the LLM considers when generating a response must appear on this whiteboard. Conversely, the LLM can only write its response to this same space.

The metaphor helps explain several important characteristics:

Limited space: Whiteboards have a finite area. Similarly, context windows can hold only a limited amount of text. We measure this capacity in tokens—roughly equivalent to words, though the precise mapping is more complex.

Temporary storage: When you erase a whiteboard, the information disappears. When an LLM finishes processing a request, the context window contents are cleared, ready for the next interaction.

External to the core: The whiteboard exists outside the brain in the jar. Similarly, context windows sit in temporary memory outside the LLM's trained parameters. This separation explains why conversations don't alter the model itself.

Single view: The brain can only process what currently appears on the whiteboard. If earlier conversation exists, it must be rewritten on the whiteboard for the brain to consider it. The brain cannot reach back to remember—it can only read what's presently visible.

This architecture has profound implications for how we design discovery workflows. Everything you want the LLM to consider—your instructions, the documents to analyze, the questions to answer, and the conversation history—must fit within this context window simultaneously.

How ChatGPT and Other Interfaces Work

Understanding the distinction between interfaces like ChatGPT and the underlying LLMs clarifies how conversations are maintained despite the LLM's lack of memory.

The Interface Layer

ChatGPT is not GPT. This seemingly obvious point bears emphasizing because many users conflate the two. ChatGPT is a software application that provides a browser-based interface for interacting with GPT. The distinction matters because the interface handles several functions beyond what the LLM itself does:

Conversation storage: ChatGPT saves your complete conversation history in its database. You can close your browser, return days later, and continue exactly where you left off. The LLM itself retains nothing—all memory resides in the interface.

Context window management: When you submit a new message in an existing conversation, ChatGPT retrieves the prior conversation history and sends it to GPT along with your new message. This happens automatically and invisibly. From your perspective, you're just continuing a conversation. From GPT's perspective, it's receiving a completely new request that happens to include contextual information.

Session management: The interface tracks multiple conversations, allows you to organize them, search through them, and delete them. It manages authentication, billing, and usage limits. None of this involves the LLM directly.

User experience: The interface formats responses for readability, provides convenient copying and export options, and creates the familiar chat-like interaction pattern. The LLM just generates text—the interface makes that text useful.

Other AI interfaces work similarly. Claude's web interface, Microsoft's Copilot, Google's Gemini interface, and countless specialized legal AI tools all follow this pattern: an interface layer



manages conversation state and user experience while routing requests to underlying LLMs.

How a Conversation Actually Works

When you start a conversation by typing a question, here's what happens:

- 1. You submit a prompt:** Your message goes to the interface (ChatGPT, Claude, etc.).
- 2. The interface prepares the context window:** The interface takes your prompt and writes it to the context window, which will be sent to the LLM.
- 3. The LLM processes:** GPT (or Claude, or Gemini) reads the context window, processes the information, and generates a response. It writes this response back to the context window.
- 4. The interface retrieves and displays:** The interface reads the LLM's response from the context window, displays it to you, and stores both your prompt and the response in its database.
- 5. Context window is cleared:** Once the interface retrieves the response, the context window is cleared, ready for the next request.

Now suppose you want to continue the conversation. You type a follow-up question. Here's what happens:

- 1. You submit a follow-up:** Your new message goes to the interface.
- 2. The interface reconstructs context:** The interface retrieves your entire conversation history from its database—all previous messages from both you and the AI.
- 3. The interface prepares the context window:** It writes the full conversation history plus your new message onto the context window.
- 4. The LLM processes as if new:** GPT receives what appears to it as a completely new request. It has no memory of responding to your earlier questions. But because the context window contains the full conversation history, it can generate a response that appropriately continues the dialogue.
- 5. The cycle repeats:** Response generated, displayed, stored, context window cleared.

This explains how LLMs can “remember” conversations despite having no memory. They don't remember—they're simply provided with the complete conversation history every time they generate a response. The illusion of memory comes entirely from the interface's storage and context window management.

Important Implications

This architecture has several implications worth noting:

Each request is independent: From the LLM’s perspective, every request is brand new. It doesn’t gradually learn about your case or remember your preferences (though we’ll discuss the recent “memory” feature later). Each response is generated fresh from the context window contents.

Different compute nodes: Your second request in a conversation might be handled by a completely different instance of GPT running on different hardware. The interface routes requests to available compute resources. This works seamlessly because everything needed sits in the context window, not in any persistent state.

Interface differences matter: Different interfaces to the same LLM can produce different user experiences. One might summarize long conversations before sending them to the LLM. Another might send complete verbatim history. These choices affect both functionality and cost.

Data persistence: While the LLM retains nothing, the interface stores your complete conversation history. For security and confidentiality, understanding where and how interfaces store data matters greatly. We explore this in Chapter 4.

Context Window Evolution: From 4k to 1M+ Tokens

The size of context windows has expanded dramatically since ChatGPT’s 2022 release. Understanding this evolution helps contextualize current capabilities and anticipate how growing windows might affect discovery workflows.

The Early Days: 4k Tokens

When GPT-3.5 first powered ChatGPT, it offered a 4,096-token context window. This translates to roughly 3,000 words or about six single-spaced pages. While revolutionary for its time, this limitation significantly constrained practical applications.

Consider discovery implications: You couldn’t ask the system to analyze even a moderately long deposition transcript. A typical single-day deposition might run 150-300 pages, far exceeding the available space. Contract review faced similar constraints. Most commercial contracts exceed a few pages, and comparative analysis across multiple contracts was impossible.

The conversation history consumed context window space alongside any documents you wanted analyzed. A lengthy back-and-forth would exhaust the available capacity, forcing the interface to truncate early conversation turns. The LLM would begin “forgetting” the start of your discussion.

Rapid Expansion: 8k, 16k, 32k

Recognition of these limitations drove rapid development. Within months, context windows expanded to 8k, then 16k, then 32k tokens. Each doubling enabled new use cases.

At 32k tokens (approximately 24,000 words), you could analyze substantial documents—perhaps a short deposition or several contracts simultaneously. This opened practical applications in legal review that simply weren’t feasible at 4k.

The Jump to 128k and Beyond

In late 2023 and early 2024, another leap occurred. GPT-4 Turbo arrived with a 128k token context window. Anthropic released Claude 3 with 200k tokens. These capacities changed the game substantially.

At 128k tokens (roughly 96,000 words), you can analyze:

- Most single-day deposition transcripts completely
- Several average-length contracts simultaneously
- Lengthy reports or legal memoranda
- Substantial portions of document productions

At 200k tokens, capacity grows to approximately 150,000 words—enough for multiple transcripts, extensive contract portfolios, or comprehensive document collections.

Both Anthropic and Google announced even larger experimental windows, approaching or exceeding one million tokens. Anthropic demonstrated Claude analyzing “The Great Gatsby” complete with scripts from all nine Star Wars films—pure marketing, but illustrative of scale. Google’s Gemini similarly claimed a million-token capacity.

Current State: 128k-200k as Standard

As of this writing, most advanced LLMs offer context windows in the 128k-200k token range as their standard configuration. This represents roughly 100,000-150,000 words, or 200-300 pages of text. For many discovery tasks, this proves sufficient to analyze substantial documents or collections.

The million-token windows remain somewhat experimental. They’re available in limited configurations, often with substantial cost premiums, and questions persist about how effectively LLMs use such extensive context.

The “Lost in the Middle” Problem

Larger context windows solve some problems but create others. Research has revealed a concerning pattern: LLMs often struggle to use information effectively when it appears in the middle of very long contexts.

What the Research Shows

Studies testing LLM performance on question-answering tasks across varying context lengths found consistent patterns. When relevant information appears at the beginning or end of the context window, models retrieve and use it effectively. But when the same information appears in the middle—particularly in very long contexts—performance degrades significantly.

Imagine asking an LLM to find a specific contractual provision buried in page 47 of a 100-page context window containing multiple contracts. Research suggests the model might miss it, even though the information technically sits within its context window. The same provision appearing in the first few pages or final few pages stands much higher chance of being noticed

and used.

This “lost in the middle” phenomenon appears to worsen as context windows grow. Paradoxically, adding more capacity sometimes reduces effective performance if critical information gets buried in the expanded middle section.

Why This Happens

The technical explanation involves how transformer architectures process sequence information and how attention mechanisms distribute across very long contexts. Without diving into neural network details, the essential point is that LLMs don’t maintain truly uniform awareness across their entire context window. Certain positions—beginnings and ends especially—receive disproportionate attention.

Think of it as analogous to human reading comprehension. If you skim a very long document quickly, you’ll likely remember points from the introduction and conclusion better than random facts from the middle chapters. LLMs exhibit similar patterns, though for different underlying reasons.

Implications for Discovery Work

This limitation has practical consequences for designing discovery workflows:

Document placement matters: When analyzing multiple documents, consider placing the most important ones at the beginning or end of the context window rather than in the middle.

Explicit references help: Rather than asking “What does the discovery plan say about privilege review?” and hoping the LLM finds the relevant section buried mid-document, you might say “Review the section on page 12 discussing privilege review and summarize it.” Explicit direction helps overcome position bias.

Chunking strategies remain relevant: Despite larger windows, breaking long documents into focused chunks and analyzing them separately sometimes produces better results than feeding everything to a massive context window at once.

Verification becomes critical: The lost-in-the-middle problem means you cannot assume the LLM considered all information in a large context window. Verification of key findings against source documents remains essential.

Window size optimization: Bigger isn’t always better. A focused analysis of carefully selected documents in a moderate context window might outperform a kitchen-sink approach that dumps everything into a million-token window.

Practical Implications for Discovery Work

Understanding context windows helps you design more effective discovery workflows. Here we explore when different approaches work best and how to work within context window constraints.

When Larger Windows Help

Single-document analysis: When you need comprehensive analysis of one substantial document—a lengthy contract, complete deposition, or detailed report—larger context windows allow the LLM to consider the entire document at once. This preserves context that might be lost if you analyzed the document in chunks.

Comparative analysis: Comparing provisions across multiple contracts, finding contradictions between witness testimonies, or identifying patterns across documents all benefit from having everything visible simultaneously in the context window.

Conversation continuity: In complex investigations requiring extensive back-and-forth with the AI, larger windows allow more conversation history to remain accessible without truncation. This maintains continuity across dozens of exchanges.

Reduced API calls: Analyzing everything in one large context window rather than multiple smaller ones reduces the number of API calls needed, potentially lowering costs and simplifying workflow architecture.

When RAG/Chunking Strategies Work Better

Very large document collections: No context window, no matter how large, can hold millions of discovery documents. When analyzing substantial document populations, retrieval-based strategies remain essential. The system searches to find relevant documents, then analyzes those specifically rather than attempting to cram everything into the context window.

Precision over breadth: Sometimes you don't need everything in view simultaneously. If you're searching for specific types of clauses across a contract portfolio, retrieving and analyzing just those sections might prove more accurate than presenting all contracts together and asking the LLM to find relevant provisions.

Cost management: Larger context windows cost more to use. Each token sent to the LLM incurs charges. For large-scale review, the costs of repeatedly sending huge contexts can become prohibitive. Targeted retrieval followed by focused analysis often proves more economical.

Avoiding lost-in-the-middle: As discussed earlier, extremely large contexts can reduce effectiveness. Strategic chunking that presents information in digestible pieces sometimes produces better results than maximum-size windows.

Iterative investigation: Discovery investigations often proceed iteratively. You find some information, which raises new questions, leading to new searches. This natural workflow pattern doesn't require everything in view at once. Instead, you benefit from targeted retrieval in each phase, analyzing focused sets of relevant documents as you progress.

Hybrid Approaches

Many effective discovery workflows combine approaches. Consider this pattern:

- 1. Initial broad search:** Use semantic search or AI classification to identify potentially rel-

evant documents from a large collection.

2. **Cluster and prioritize:** Group similar documents and rank by likely relevance.
3. **Focused analysis:** Analyze high-priority document clusters using appropriately-sized context windows—large enough for meaningful synthesis, small enough to avoid lost-in-the-middle problems.
4. **Iterative refinement:** Based on findings, conduct targeted searches for specific information, analyzing results in focused contexts.
5. **Synthesis:** Combine insights from multiple focused analyses into comprehensive understanding.

This hybrid approach leverages both retrieval methods (for finding documents) and context window capacity (for analyzing them), avoiding the weaknesses of either extreme.

Cost Considerations with Large Contexts

Context window size directly affects cost. AI providers typically charge based on tokens processed, with separate rates for input tokens (what you send) and output tokens (what the AI generates). Larger contexts mean more input tokens, increasing costs.

Consider the economics:

Small context (8k tokens): Analyzing a short document costs perhaps a few cents. Conducting thousands of such analyses across a document collection remains affordable.

Medium context (128k tokens): Analyzing a substantial document or small collection might cost several dollars per analysis. Scaling this across thousands of analyses becomes expensive but potentially justifiable for complex matters.

Large context (1M tokens): A single analysis might cost tens of dollars. This makes sense for critical documents requiring comprehensive analysis but proves impractical for routine review across large populations.

Conversation history: Remember that every message in an ongoing conversation includes all prior messages as context. A conversation with 50 exchanges might accumulate substantial token costs as the context window fills with history. Some interfaces implement strategies to manage this—summarizing old conversation turns, pruning less relevant history, or using separate sessions for distinct topics.

For discovery budgets, these cost structures suggest tailoring context window usage to task importance. Reserve large context windows for critical analyses where comprehensive synthesis matters. Use smaller, focused contexts for routine review tasks. Design workflows that minimize redundant token usage—don't resend the same documents repeatedly if you can cache analyses.

Why LLMs Have No Memory (And Why That Matters)

We've explained that LLMs don't remember conversations—the interface handles that through context window management. But this characteristic deserves deeper exploration because it has security and practical implications.

Technical Reality: Fixed Parameters

When an LLM completes training, its neural network parameters are set. These parameters—essentially the “weights” that determine how the model processes input and generates output—remain fixed during inference (the process of generating responses to prompts).

Your prompts don't modify these parameters. The model reads your input, processes it through its fixed neural network, generates output, and returns to exactly the same state afterward. It's genuinely stateless between requests.

This differs fundamentally from how humans learn through conversation. When you explain something to a colleague, they retain that information. Their understanding evolves. An LLM doesn't. Each interaction is independent, processed through the identical neural network structure.

The “Memory” Feature Caveat

Recent versions of ChatGPT and some other interfaces offer a “memory” feature that might seem to contradict what we've said. This feature allows the interface to maintain persistent facts about you across conversations—your preferences, your role, recurring topics you discuss.

However, this “memory” isn't LLM learning. Instead, the interface maintains a separate database of facts about you. When you start a new conversation, the interface includes relevant facts from this database in the context window sent to the LLM. From the LLM's perspective, these are just more instructions in the current context window, not actual memories.

This distinction matters because:

The LLM still knows nothing about you: Different users querying the same LLM with identical prompts get identical responses (accounting for randomness in text generation). Your history doesn't affect the underlying model.

The interface controls memory: Interface providers decide what to save, how long to retain it, and when to include it in contexts. The LLM has no agency in this process.

Security implications remain unchanged: The LLM cannot leak your information to other users because it never stored that information in the first place. We explore this crucial security point in Chapter 4.

Practical Implications

Understanding that LLMs lack memory helps you use them effectively:

Provide complete context: Don't assume the LLM remembers earlier statements unless they appear in the current conversation. If you discussed a case yesterday in one conversation and start a new conversation today, you must reintroduce the case context.

Conversation organization: Since conversations maintain independent context, you might deliberately use separate conversations for separate matters. This prevents unrelated context from consuming token space and keeps discussions focused.

Session management: Some discovery workflows benefit from long-running sessions where extensive context accumulates. Others work better with frequent fresh starts that clear context and begin from clean states. Understanding the stateless nature helps you choose appropriately.

Reproducibility: Because the LLM has no state beyond what's in the current context window, providing identical inputs produces consistent outputs (within the bounds of generation randomness). This supports validation and verification workflows.

Training doesn't happen during use: You cannot "teach" an LLM through conversation. Fine-tuning requires returning to the training process with new data. For legal applications, this means domain knowledge must come from careful prompting, retrieval of relevant information, or using models specifically fine-tuned for legal work before deployment.

Looking Ahead

This chapter examined how LLMs interact with users through context windows managed by interface layers. We explored the evolution of context window sizes, the lost-in-the-middle problem, and practical implications for discovery workflows. Understanding these mechanisms allows you to design more effective applications of AI in legal work.

But even with this knowledge, a critical concern remains: accuracy. LLMs can generate impressively fluent text that's factually wrong, a phenomenon called hallucination. For legal professionals, where precision and reliability are paramount, this represents perhaps the most serious challenge in AI adoption.

Chapter 3 addresses hallucinations comprehensively—what they are, why they occur, how to detect them, and most importantly, how to structure workflows that minimize their risk while preserving the efficiency gains AI offers. Understanding hallucinations isn't optional for responsible AI use in discovery. It's foundational to maintaining professional standards while leveraging these powerful tools.

03

Hallucinations and Accuracy

—

CHAPTER THREE

Hallucinations and Accuracy

The most serious concern about using Large Language Models in legal work isn't their cost, complexity, or learning curve. It's their tendency to generate plausible-sounding content that is simply wrong. This phenomenon, called "hallucination," represents the fundamental challenge that legal professionals must understand and address before deploying AI systems in practice.

Unlike human mistakes, which typically involve misunderstanding, oversight, or carelessness, LLM hallucinations often appear confident and detailed. The system doesn't indicate uncertainty or flag potential errors. It generates fabricated case citations, non-existent statutes, or invented facts with the same fluency it uses for accurate information. For legal professionals accustomed to associates who acknowledge when they don't know something, this behavior can be dangerously misleading.

This chapter examines hallucinations comprehensively: what they are, why they occur, how to recognize them, and most importantly, how to structure workflows that preserve AI's efficiency benefits while minimizing accuracy risks. Understanding hallucinations isn't merely advisable for legal AI use—it's essential to maintaining professional competence and avoiding the career-damaging mistakes several lawyers have already made.

What Are Hallucinations?

A hallucination occurs when an LLM confidently generates detailed information that is factually incorrect, fabricated, or cannot be verified in its training data or provided context. The term "hallucination" captures the phenomenon well: the system appears to perceive information that doesn't exist, much as a hallucinating person might see things that aren't there.

Key Characteristics of Hallucinations

Plausibility: Hallucinated content isn't random nonsense. It follows grammatical rules, fits contextual expectations, and often incorporates real elements. A hallucinated case citation includes proper formatting, realistic party names, and appropriate court identifiers. This plausibility makes hallucinations dangerous—they don't trigger immediate skepticism.

Confidence: LLMs generate hallucinated content with the same apparent certainty they use for accurate information. There's no hedging, no indication of uncertainty, no disclaimer that the information might be fabricated. The system presents invented facts as definitively as verified ones.



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

Detail: Hallucinations often include specific, concrete details that enhance their apparent credibility. Rather than vaguely mentioning “a federal case,” the system might cite “Johnson v. Peterson, 847 F.3d 592 (9th Cir. 2019),” complete with plausible docket numbers, judge names, and holding descriptions—none of which exist.

Internal consistency: Hallucinated information often maintains internal logical consistency. If the system invents a case, subsequent references to that case will remain consistent with the initial fabrication. This consistency across multiple references reinforces the false impression of accuracy.

Why Hallucinations Occur

Understanding why LLMs hallucinate requires remembering what they fundamentally do: predict probable next words in sequences based on training patterns. They aren’t searching databases of verified facts. They’re generating statistically plausible continuations of text.

When you ask an LLM about a real case it encountered during training, it generates text based on patterns it learned from thousands of case discussions. The output might accurately reflect that training data. But when you ask about something not in its training data, the system doesn’t respond “I don’t know.” Instead, it generates plausible-sounding text following the patterns it learned, effectively inventing information that fits the expected format.

Consider this analogy: If you ask someone to describe a Supreme Court case they’ve never heard of, an honest person says they don’t know. But imagine someone trained to always provide answers, never admit ignorance, and construct responses from general knowledge of how legal writing works. They might generate a plausible-sounding description that follows all the right patterns—procedural posture, legal issues, holding, reasoning—while inventing every substantive detail. That’s essentially what LLMs do when hallucinating.

The training process teaches LLMs to generate fluent, contextually appropriate text. It doesn’t teach them to distinguish between “I learned this specific fact during training” and “I can generate plausible text fitting this context.” The model has no internal mechanism for evaluating whether generated content corresponds to reality. It only evaluates whether the text flows naturally and fits the pattern.

Fine-tuning through human feedback reduces hallucination frequency by teaching models to be more careful, acknowledge uncertainty, and provide better-grounded responses. But fine-tuning doesn’t eliminate the fundamental tendency. The model architecture still generates text through statistical prediction, not fact verification.

Hallucination vs. Mistaken Inference

It’s important to distinguish hallucinations from legitimate analytical errors. LLMs can make mistaken inferences—drawing incorrect conclusions from ambiguous text, misinterpreting nuanced arguments, or applying faulty logic to genuine information. These are analytical failures, similar to mistakes human lawyers make when reasoning from complex sources.

Hallucinations are different. They involve fabricating information wholesale. The distinction matters because different mitigation strategies apply:

Mistaken inferences require better prompting, clearer instructions, human review of reasoning, and verification of analytical conclusions. You might address them by providing more context, asking the system to explain its reasoning, or requesting alternative interpretations.

Hallucinations require ensuring the system bases outputs on verified source material rather than generating from training memory. You address them through retrieval-based architectures, mandatory citation requirements, and systematic verification of factual claims against source documents.

A system that reads a contract and misinterprets an ambiguous clause made an analytical error. A system that invents contract provisions that don't exist in the source document hallucinated. The first is a reasoning problem; the second is a reliability problem.

The Fake Cases Crisis: Lessons from Recent Sanctions

The legal profession's introduction to AI hallucinations came through a series of high-profile cases where lawyers submitted court filings citing non-existent cases generated by ChatGPT. These incidents provide sobering lessons about hallucination risks and the consequences of insufficient verification.

The Mata v. Avianca Case

In May 2023, lawyers Steven A. Schwartz and Peter LoDuca filed a brief in federal court in New York opposing a motion to dismiss. The brief cited multiple cases supporting their arguments, including *Vargas v. China Southern Airlines Co.*, *Shaboon v. Egyptair*, and *Petersen v. Iran Air*. The citations appeared properly formatted with case numbers, court identifiers, and holding descriptions.

Opposing counsel couldn't locate these cases in legal databases. When questioned, the lawyers provided additional details—quotes from the purported decisions, internal citations to other cases, and explanations of the holdings. All fabricated.

The lawyers had used ChatGPT to conduct legal research. When ChatGPT generated the fake citations, they didn't verify them in actual legal databases. When opposing counsel questioned the cases, they returned to ChatGPT asking for confirmation. ChatGPT, unable to acknowledge its earlier hallucinations, provided additional fabricated details supporting the fake cases' existence.

The court was not amused. Judge Kevin Castel sanctioned both lawyers, fining each \$5,000 and referring them to grievance committees for potential disciplinary action. In his opinion, Judge Castel emphasized that while using AI for legal research wasn't improper per se, lawyers maintain responsibility for ensuring accuracy of their filings. The opinion noted that even a cursory verification effort would have revealed the cases didn't exist.

The Park v. Kim Case

Shortly after the Mata sanctions, attorney Jae Lee faced similar problems in a federal appeal. Lee cited a non-existent state court decision in briefing to the Second Circuit Court of Appeals. The citation came from ChatGPT, and again, basic verification would have revealed the fabri-

cation.

The Second Circuit referred Lee to a grievance panel for potential sanctions, noting the concerning pattern of lawyers submitting AI-generated fake cases without verification.

Additional Incidents

These weren't isolated incidents. Several other cases emerged involving fake citations from AI systems:

In Colorado, an attorney submitted a brief citing cases that didn't exist, later explaining he had used an AI tool without proper verification.

In Texas, another lawyer faced sanctions for similar conduct, with the court emphasizing that technological tools don't diminish attorneys' responsibility for accuracy.

The pattern across these cases is remarkably consistent: lawyers used ChatGPT or similar tools for legal research, received hallucinated cases, failed to verify them in legal databases, submitted them to courts, and faced professional consequences when the fabrications were discovered.

What Went Wrong

These incidents reveal multiple failures in understanding and using AI:

Misunderstanding AI capabilities: The lawyers apparently believed ChatGPT functioned like a legal research database—searching and retrieving actual cases. In reality, it was generating plausible-sounding text based on patterns, not retrieving verified citations.

Absence of verification: In traditional legal practice, if an associate provides case citations, prudent partners verify key cases before filing. These lawyers skipped verification, apparently trusting AI output implicitly.

Compounding errors: When questioned, several lawyers returned to ChatGPT for confirmation rather than checking actual legal databases. ChatGPT, unable to admit error, provided additional fabricated details, deepening the problem.

Inadequate understanding of hallucinations: The lawyers likely didn't understand that AI systems could generate confident, detailed, completely fabricated legal citations. They learned this expensive lesson publicly.

Professional responsibility failures: The sanctions opinions emphasized that lawyers cannot delegate their professional duty of competence to technology. Using AI tools doesn't excuse submitting false information to courts.

Lessons for Discovery Professionals

While these cases involved litigation briefs rather than discovery, the lessons apply directly to discovery work:

Verification is mandatory: AI-generated information requires verification against source documents, regardless of how confident or detailed the output appears.

Understand tool limitations: Know what your AI tools actually do. Are they searching databases and retrieving documents? Or generating text based on patterns? The distinction is critical.

Design verification into workflows: Don't treat verification as optional or reserve it for suspicious outputs. Build systematic verification into every AI-assisted process.

Human responsibility remains: AI assistance doesn't diminish professional obligations. Discovery professionals remain responsible for accuracy, regardless of what tools they employ.

Training is essential: Organizations deploying AI tools must ensure users understand both capabilities and limitations. The sanctioned lawyers might have avoided their predicament with proper training about hallucinations.

Hallucination Risk Factors

Not all AI applications present equal hallucination risk. Understanding factors that increase or decrease risk helps you assess when to use AI, what verification strategies to employ, and where extra caution is warranted.

High-Risk Scenarios

Relying on training memory: When you ask LLMs questions that require them to recall information from training data—"What does the Jones Act say about seaman status?" or "Summarize the holding in *Iqbal*"—hallucination risk is high. The system might have seen relevant information during training, or it might generate plausible-sounding responses based on general patterns.

Generating citations: Any task requiring case citations, statutory references, or specific source attributions presents substantial hallucination risk. LLMs can easily generate properly formatted citations to non-existent sources.

Precise factual claims: Questions requiring specific facts, dates, numbers, or technical details invite hallucinations. "When was the EPA regulation adopted?" might receive a confidently stated but fabricated year.

Obscure or specialized topics: The less common a topic in training data, the more likely hallucinations become. Asking about a rarely cited case, an obscure regulation, or a novel legal theory increases risk that the system will generate plausible-sounding fabrications.

Complex multi-step reasoning: Tasks requiring the system to combine multiple facts, perform calculations, or execute multi-step logic without source documents present elevated hallucination risk. Each reasoning step introduces additional opportunity for error or fabrication.

Lower-Risk Scenarios

Document summarization: When you provide source documents and ask the system to summarize them, hallucination risk decreases substantially. The system bases its response on the provided text rather than generating from training memory.

Pattern identification: Tasks like identifying document types, classifying content, or recognizing themes in provided documents present lower hallucination risk because they operate on visible input rather than recalled information.

Draft generation: Using AI to draft initial versions of documents—demand letters, deposition outlines, privilege logs—carries lower hallucination risk in terms of factual fabrication. The drafts may need substantial revision for quality, but they typically don't invent facts from nothing.

Question answering from context: When you provide documents and ask specific questions answered in those documents, hallucination risk is lower. The system can extract information from visible context rather than generating from memory.

Comparative analysis: Asking the system to compare provisions across provided contracts, identify differences between documents, or analyze patterns in visible texts reduces hallucination risk. The comparison operates on provided source material.

Mitigation Through Design

Understanding risk factors allows designing workflows that minimize exposure:

Emphasize extraction over generation: Tasks that extract information from provided documents carry less risk than tasks requiring the system to generate information from training memory.

Require citations: Mandate that every factual claim include citations to source documents. This allows verification and reduces pure fabrication.

Limit scope to provided materials: Instruct systems explicitly to base responses only on provided documents, not on training knowledge. This dramatically reduces hallucination risk. Implement verification steps: Build verification into workflows rather than treating it as optional. Have humans or secondary AI systems check key facts against source documents.

Use specialized tools appropriately: For tasks like legal research requiring verified citations, use databases designed for that purpose rather than general-purpose LLMs.

Mitigation Strategies: RAG, REG, and Verification

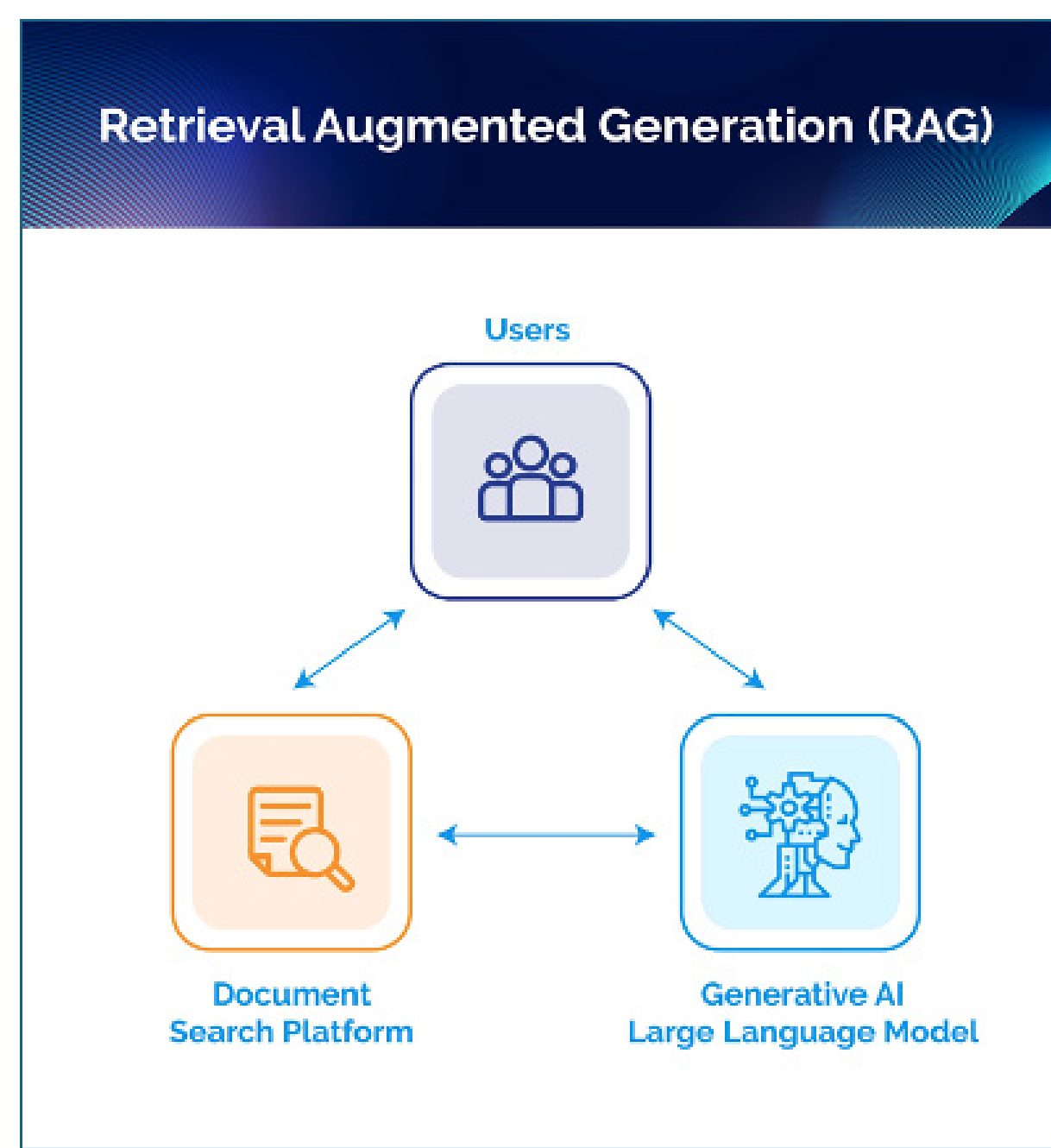
The discovery profession has developed several approaches to minimize hallucination while preserving AI's efficiency benefits. Understanding these strategies and their appropriate applications is essential for responsible AI deployment

Retrieval Augmented Generation (RAG)

RAG represents a fundamental shift in how we use LLMs. Rather than asking the system to answer questions from its training memory, RAG systems first search document collections to find relevant passages, then provide those passages to the LLM as context for generating responses.

How RAG Works:

- 1. User query:** You submit a question or request.
- 2. Document retrieval:** The system searches a document collection using semantic search, keyword matching, or hybrid approaches to find relevant passages.
- 3. Context assembly:** Retrieved passages are formatted and placed in the LLM's context window along with your query.
- 4. Prompted response:** The LLM receives explicit instructions to base its response on the provided passages rather than its training memory.
- 5. Generated output:** The system generates a response grounded in the retrieved documents.



A Typical RAG Architecture

Benefits of RAG:

Reduced hallucination: By constraining the LLM to provided source material, RAG dramatically reduces fabrication risk. The system can't invent information not present in retrieved documents.

Source attribution: RAG systems can cite specific retrieved passages, allowing verification of outputs against source material.

Current information: RAG overcomes training cutoff limitations. The document collection can include material from after the LLM's training date, making the system capable of addressing current information.

Domain specificity: RAG allows you to create domain-specific AI applications by controlling what documents the system searches. A contracts AI might search only your organization's contract repository, ensuring responses reflect your specific agreements rather than generic training data.

Challenges with RAG:

Retrieval quality matters: If the search step fails to find relevant documents, the LLM lacks information needed to answer accurately. Poor retrieval undermines the entire approach.

Context window limits: Only top-ranked retrieved documents fit in the context window. Important information in lower-ranked results might be missed.

Complexity: RAG systems require more infrastructure than simple LLM queries—document collections, search capabilities, orchestration logic, and careful prompt engineering.

Retrieval Exclusive Generation (REG)

Building on RAG principles, we developed an even more stringent approach called Retrieval Exclusive Generation. REG emphasizes that systems should use exclusively retrieved information, completely eliminating reliance on the LLM's pre-trained knowledge.

How REG Differs from RAG:

Explicit exclusivity instructions: We instruct LLMs twice—once in system-level instructions and again in user prompts—to base analysis solely on provided documents, explicitly prohibiting use of training knowledge.

Mandatory source citations: Every substantive claim in the output must include direct citations to source documents with specific identifiers allowing immediate verification.

Verification protocols: REG implementations include systematic verification steps where humans or secondary systems check that cited sources actually support claimed information.

Rejection of outside knowledge: If the system cannot answer a question based on provided documents, it must acknowledge this rather than supplementing with training knowledge.

Discovery Applications:

The book's earlier examples demonstrate REG in practice. In the Bacardi trademark investigation example, the system searched a large document collection, retrieved relevant documents,

analyzed only those documents, and generated a report with citations to specific document IDs. Every claim in the report could be traced directly to source material.

Similarly, in the transcript analysis example, the system searched across deposition transcripts, retrieved relevant sections, and generated answers citing specific transcript sections and page numbers. Nothing in the report came from the LLM's general training—everything derived from the provided transcripts.

Benefits of REG:

Maximum accuracy: By eliminating training memory as a source, REG minimizes hallucination risk to near-zero for factual claims.

Complete verifiability: Every claim includes citations allowing immediate verification against source documents.

Defensibility: When questioned about AI-assisted work product, you can demonstrate that outputs derive entirely from case materials, not from AI training data of uncertain provenance.

Professional standards: REG aligns with legal professional responsibilities to base work on verified sources rather than unverifiable AI-generated content.

Citation Requirements

Regardless of whether you implement full RAG or REG, requiring citations in AI outputs provides crucial benefits:

Verification capability: Citations allow you to check whether source documents actually support claimed information. Without citations, verification requires re-searching source materials, often proving impractical.

Hallucination detection: Hallucinations often become apparent when you attempt to verify citations. The cited document doesn't exist, doesn't say what's claimed, or doesn't support the conclusion drawn.

Quality signal: Systems that provide detailed citations tend to generate more accurate outputs. The act of citing sources appears to encourage more careful generation.

Professional documentation: Citations create the audit trail needed to demonstrate that work product derives from case materials rather than AI fabrication.

Implementation approaches:

Mandatory citation prompts: Explicitly instruct the system to cite sources for every factual claim. "Provide citations to source documents for all substantive statements" should appear in your prompts.

Citation formats: Specify citation formats that allow easy verification. Document IDs, page numbers, section references, and file names all help locate source material quickly.

Citation validation: Build workflows that check cited sources. This might involve human review of key citations or automated systems that verify cited documents exist and contain claimed information.

Rejection of uncited claims: Treat any factual claim without citation as suspect. Either require the system to provide citations or remove uncited claims from outputs.

Verification Protocols

Even with RAG, REG, and citation requirements, human verification remains essential. No technical architecture eliminates the need for professional oversight.

What to verify:

Key facts: Any fact critical to your analysis, argument, or decision requires verification. Check that source documents actually support claimed information.

Citations: Verify that cited sources exist and contain the claimed content. This proves especially important for any citations that will appear in filings or reports to clients.

Numbers and dates: AI systems often struggle with numerical precision. Any dates, dollar amounts, percentages, or quantitative claims warrant careful verification.

Logical inferences: Check that conclusions drawn from source materials are logical and supported. Even if individual facts are accurate, inferences might be flawed.

Legal standards: Verify that any statements about legal requirements, standards, or precedents accurately reflect controlling law.

Verification strategies:

Sampling: For large volumes of AI-generated output, verify a statistical sample. Use findings to assess overall reliability and identify areas needing closer review.

Risk-based verification: Verify high-stakes claims more thoroughly than routine matters. Allocate verification effort according to potential impact of errors.

Cross-checking: Compare AI outputs against other sources—traditional research, human analysis, or different AI systems. Consistency across sources increases confidence.

Expert review: Have subject matter experts review AI-generated content in their areas. A tax specialist reviewing AI-generated tax analysis can spot errors general reviewers might miss.

Adversarial review: Consider how opposing counsel might challenge AI-generated work product. Verify elements that could become points of contention.

Quantifying Risk: When to Trust AI Outputs

Understanding hallucination risks helps you make informed decisions about when to rely on AI outputs with less verification versus when to treat outputs as starting points requiring extensive validation.

The screenshot displays the Alchemy AI Evaluation Site interface. The top navigation bar includes the Alchemy logo, 'Alchemy Evaluation Site', and user information 'John Tredennick'. Below this, a breadcrumb trail shows the path: Project Details > Conflicts Testim... > Macondo > BP Fault > Cement failure > Halliburton > Toolpusher > Texas City > Indemnity. The main content area is titled 'Alchemy AI Topic: Cement failure' and shows the 'Summary LLM: Gemini 2.5 Flash Lite' and 'Reporting LLM: Claude Sonnet 4.5'. The 'Topic Discussion' section includes an 'Investigation Timeline' and 'Overall Assessment and Conclusions'. The 'Primary Responsibility' section states that all cementing-related decisions were made solely by BP. The 'Causal Relationship to Blowout' section details the failure to establish a reliable cement barrier. The 'Expert Consensus' section notes that while experts disagreed on specific technical mechanisms, there was consensus that cement problems resulted from BP's operational decisions. A 'Summaries Used' section lists 94 summaries. A 'Make a request' button is at the bottom. On the right, a 'Summaries' table lists 100 of 9374 records. The first record is 'BP_00001611 §10 Critical Examination of BP's Cement Job Failures and Disagreements Over Cement Stability' with a relevance of 85%. The second record is 'BP_00001391 §17 BP Cementing Failures, Testing Shortcomings, and Expert Disagreements on Nitrogen Stability in Macondo Well' with a relevance of 92%.

Example of a REG system that requires citations to source material with links for quick verification for material statements.

Factors Affecting Reliability

Specificity of sources: Outputs based on specific provided documents are more reliable than outputs based on training memory. If the system analyzed your case documents and cited specific sections, reliability is higher than if it generated responses from general knowledge.

Citation completeness: Outputs with comprehensive citations are more reliable than those with vague references or no citations. “According to Document ID 45892, page 7, paragraph 3” is more trustworthy than “based on the contract terms.”

Consistency across sources: Claims supported by multiple independent sources are more reliable than claims appearing in single sources, especially if those sources come from parties with aligned interests.

Domain complexity: In straightforward domains with clear factual content, AI outputs may be more reliable. In complex areas requiring nuanced interpretation or specialized knowledge, expect lower reliability without expert verification.

Output type: Summaries of provided documents are generally more reliable than analytical conclusions. Extracted facts are more reliable than legal interpretations. Direct quotes are more reliable than paraphrases.

Model quality: More recent, larger, better-fine-tuned models generally produce more reliable outputs. GPT-4o or Claude Sonnet 4.5 outputs are typically more reliable than outputs from earlier or smaller models.

Creating Reliability Frameworks

Organizations should develop frameworks for assessing output reliability and determining appropriate verification levels:

High-reliability outputs (minimal verification needed):

- Summaries of provided documents with comprehensive citations
- Extracted structured data from clear sources
- Categorization of provided documents
- Timeline generation from documents with date verification

Medium-reliability outputs (moderate verification needed):

- Analytical conclusions from provided documents
- Comparative analyses across multiple sources
- Identification of themes and patterns
- Draft work product requiring professional review

Low-reliability outputs (extensive verification required):

- Anything based on training memory without source documents
- Complex legal analysis requiring specialized expertise
- Outputs intended for court filings or client deliverables
- Novel or unusual conclusions not obviously supported by sources

No-reliability outputs (treat as purely generative without evidentiary value):

- Legal research without verified citations
- Factual claims without source attribution
- Technical or specialized content outside verified sources
- Predictive or speculative content

Cost-Benefit Analysis

Verification requires time and resources. Decisions about verification intensity should weigh reliability needs against practical constraints:

For document review in discovery, you might tolerate higher error rates than for court filings. If you're identifying potentially relevant documents from millions of records, and 5% classification errors are acceptable given manual review follows, extensive verification of AI classifications might be unnecessary.

Conversely, for deposition summaries that attorneys will rely on for trial preparation, verification intensity should be high. Missed contradictions or inaccurate attributions could undermine case strategy.

The key is matching verification effort to stakes. Don't treat all AI outputs identically. Invest verification resources where errors matter most.

The Human-in-the-Loop Imperative

Throughout this chapter, we've emphasized verification, oversight, and human judgment. This reflects a fundamental principle: AI should augment human expertise, not replace it. The phrase "human-in-the-loop" captures this principle—AI generates outputs, but humans remain actively engaged in reviewing, verifying, and taking responsibility for work product.

Why Humans Remain Essential

Professional judgment: AI can analyze documents and identify patterns, but professional judgment determines strategy, assesses significance, and makes tactical decisions. These require experience, wisdom, and contextual understanding that AI systems lack.

Contextual understanding: Humans understand stakes, relationships, politics, and nuances that AI systems cannot grasp. A document might be technically non-privileged but strategically sensitive. AI sees only technical analysis.

Error detection: Humans can recognize when AI outputs don't make sense, conflict with known facts, or seem anomalous. This common-sense check catches errors that might otherwise propagate through workflows.

Ethical responsibility: Professional obligations rest on humans, not tools. Lawyers and discovery professionals cannot delegate ethical responsibility to AI systems. They remain accountable for work product quality and ethical compliance.

Adaptation and learning: Humans learn from errors, adapt strategies when approaches fail, and apply creativity to novel situations. AI systems lack this adaptive capability within deployed applications.

Effective Human-AI Collaboration

The goal isn't minimizing human involvement—it's optimizing the division of labor between humans and AI:

AI handles volume: AI excels at processing large volumes quickly. Use it to analyze thousands of documents, generating initial summaries or classifications that humans would find prohibitively time-consuming.

Humans handle judgment: Reserve human attention for decisions requiring professional judgment, strategic thinking, or ethical considerations. Don't waste human capacity on tasks AI handles well.

AI generates drafts: Use AI to create initial versions of work product—summaries, outlines, analyses—that humans then refine, verify, and finalize.

Humans verify key elements: Focus human verification on critical facts, important conclu-

sions, and high-stakes elements where errors would be most consequential.

Iterative refinement: Establish workflows where AI generates outputs, humans review and provide feedback, AI refines based on that feedback, and the cycle continues until quality standards are met.

Looking Ahead

This chapter examined the most serious challenge in legal AI adoption—the tendency of LLMs to generate plausible but fabricated content. We explored why hallucinations occur, what makes them dangerous, and how to structure workflows that minimize their impact while preserving efficiency benefits.

The strategies we discussed—RAG and REG architectures, mandatory citations, systematic verification, and human-in-the-loop workflows—represent current best practices. But they don't eliminate all risks. Responsible AI use in discovery requires ongoing vigilance, healthy skepticism toward AI outputs, and commitment to verification.

Understanding hallucinations prepares you to use AI effectively. But another concern often dominates initial discussions about AI in legal practice: security and confidentiality. Can you safely send client information to AI systems? Do you risk waiving privilege or breaching confidentiality obligations? These questions deserve thorough examination.

Chapter 4 addresses data security, privacy, and confidentiality comprehensively. We explore where your data goes when you use AI systems, what protections exist, what risks you should understand, and how to assess whether AI use aligns with your professional obligations to protect client information. Security concerns are legitimate, but as you'll see, they may be more manageable than many legal professionals initially assume.

04

Data Security, Privacy, and Confidentiality

CHAPTER FOUR

Data Security, Privacy, and Confidentiality

When ChatGPT burst onto the scene in late 2022, one question dominated discussions among legal professionals: “Can I send client data to this system without breaching confidentiality or waiving privilege?” The concern was understandable. Lawyers face strict ethical obligations to protect client information, and the prospect of transmitting sensitive documents to cloud-based AI systems raised legitimate questions about security, data retention, and inadvertent disclosure.

These concerns haven’t disappeared, but our understanding has matured considerably. This chapter addresses data security and confidentiality comprehensively, examining how AI systems handle data, what protections exist, what risks remain, and how to assess whether AI use aligns with professional obligations.

The conclusion many discovery professionals have reached may surprise those still operating from initial fears: using commercial AI services through proper channels presents security and confidentiality risks comparable to other cloud services legal professionals routinely employ, such as email, document management systems, and ediscovery platforms. In some respects, AI systems may be more secure than traditional alternatives.

Understanding AI Data Flow

Before assessing security, we must understand what happens to data you send to AI systems. The answer depends entirely on which service you use and how you access it.

The Consumer vs. Commercial Divide

This distinction is fundamental and often misunderstood. Consumer AI services and commercial AI services handle data very differently, with profound implications for legal use.

Consumer Services:

Consumer versions of AI platforms—the free ChatGPT website, free Claude interface, or free Gemini access—typically reserve rights to use your inputs for various purposes. When OpenAI first released ChatGPT as a free public beta, its terms explicitly allowed using conversation data to improve the service and train future models.

For AI companies, these free offerings serve multiple purposes. They allow widespread testing and feedback collection. They generate publicity and user adoption. And critically, they provide data that helps improve models through the fine-tuning process we discussed in Chapter 1.

The trade-off seemed clear: users received free access to powerful technology, and in exchange, their inputs might be used to improve the service. For casual personal use, many found this acceptable. For professional legal work involving confidential client information, it

obviously wasn't.

Commercial Services:

Recognizing that many potential customers couldn't accept consumer terms, AI providers developed commercial offerings with fundamentally different data handling policies. These commercial services, accessed through paid subscriptions or API agreements, include strict contractual provisions prohibiting use of customer data for training, improvement, or any purpose beyond processing the specific user request.

The distinction isn't merely one of cost. It's a fundamental difference in the relationship between user and provider. Consumer services involve data exchange—you provide data that helps improve the service. Commercial services are pure service provision—you pay money, and the provider processes your data according to strict limitations.

Where Your Data Goes

When you submit information to a commercial AI service, here's the actual data flow:

- 1. Transmission:** Your prompt travels from your device to the AI provider's servers through encrypted connections (typically TLS/SSL encryption, the same used for banking and email). The data is encrypted in transit, making interception during transmission extremely difficult.
- 2. Processing:** The AI provider's servers receive your encrypted data, decrypt it in their secure environment, and route it to an available LLM instance. This might involve load balancing across data centers, but the data remains within the provider's secure infrastructure.
- 3. Context Window:** Your prompt is placed in the context window (the "whiteboard" from Chapter 2) along with any conversation history or additional context. The LLM reads this information, processes it through its neural network, and generates a response.
- 4. Response Generation:** The generated response is written back to the context window. The interface layer retrieves this response to return to you.
- 5. Temporary Storage:** Throughout this process, your data exists in temporary memory—the context window and associated processing buffers. This is volatile storage that's cleared once processing completes.
- 6. Context Window Clearing:** After the response is returned, the context window is cleared. The LLM instance returns to its baseline state, retaining no information about your request.
- 7. Return Transmission:** Your response travels back to your device through encrypted connections.
- 8. Interface Storage:** If you're using an interface like ChatGPT or Claude's web application, your conversation is stored in the interface provider's database so you can return to it later. This storage is separate from the LLM itself and subject to the interface's data re-

tention policies.

What Doesn't Happen

Equally important is understanding what doesn't occur with commercial AI services:

The LLM doesn't learn from your data: Your prompts don't modify the model's neural network parameters. The model processes your request and returns to exactly the same state afterward. This is the "brain in a jar" characteristic from Chapter 2.

Your data doesn't enter training: Commercial agreements explicitly prohibit using customer data for training future models. Your confidential documents won't appear in the next version of GPT or Claude.

Other users can't access your data: Each request is processed independently. User A's prompts never appear in User B's context window. The LLM has no mechanism to leak information between users because it doesn't retain information after processing.

Data doesn't persist indefinitely in processing systems: Once your request is completed and the response returned, the processing systems clear your data from temporary memory. It doesn't sit in databases waiting for potential breach.

Providers don't review your prompts (typically): Commercial agreements usually specify that providers won't access or review customer data except in limited circumstances like security monitoring, abuse prevention, or when required by law.

Commercial Agreements and Data Protection

The contractual terms governing commercial AI services provide the framework for data protection. Understanding these agreements helps assess whether using specific services aligns with professional obligations.

Standard Commercial Provisions

Major AI providers offer commercial terms that include several key provisions:

Data Non-Use Clauses:

These provisions explicitly state that the provider will not use customer data for training, improvement, or any purpose other than providing the service. Here's Microsoft's language for Azure OpenAI Service as a representative example:

"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”

OpenAI’s own commercial terms include similar provisions:

“We do not use Content that you provide to or receive from our API (‘API Content’) to develop or improve our Services.”

Anthropic, Google, and other providers include comparable language in their commercial agreements.

Data Retention Limits:

Commercial agreements typically specify how long customer data is retained and in what circumstances. Many specify:

- **Processing data:** Retained only as long as necessary to complete the request (typically seconds or minutes)
- **Logs:** May be retained for limited periods for security and operational purposes (30-90 days commonly)
- **Conversation history:** Retained in interface storage as long as user maintains their account, but deletable by user
- **Backups:** May exist in backup systems for limited periods but subject to same use restrictions

Zero Data Retention Options:

Some providers offer zero data retention configurations where even logs are minimized or eliminated. These configurations provide maximum security for highly sensitive applications, though they may limit troubleshooting capabilities if issues arise.

Encryption Requirements:

Commercial agreements specify encryption standards for data in transit and at rest. Industry-standard encryption (AES-256 for data at rest, TLS 1.2+ for transmission) is typical.

Compliance Certifications:

Enterprise AI providers typically maintain compliance with various security and privacy frameworks:

- SOC 2 Type II certification
- ISO 27001 certification
- GDPR compliance for European data
- HIPAA compliance capabilities for healthcare data
- Industry-specific certifications as applicable

Data Location and Sovereignty:

Some commercial agreements allow specifying where data is processed and stored geograph-

ically. This matters for organizations with data residency requirements or concerns about specific jurisdictions.

Subprocessor Disclosure:

Commercial agreements typically disclose what subprocessors (vendors the AI provider uses) have access to customer data and under what terms. This transparency allows assessment of the full security chain.

API vs. Web Interface

How you access AI services affects data handling:

API Access:

API (Application Programming Interface) connections allow your software to communicate directly with AI services without human use of web interfaces. APIs typically offer:

- **Programmatic control:** Your code controls exactly what data is sent and how
- **No conversation storage:** Unless you explicitly implement it, APIs don't store conversation history
- **Enterprise security:** API access often includes additional security options like private endpoints, enhanced authentication, and audit logging
- **Clear contractual terms:** API agreements make data handling explicit and enforceable

For production discovery workflows processing sensitive documents, API access through commercial agreements represents best practice.

Web Interfaces:

Web-based interfaces like ChatGPT's website or Claude's chat interface offer convenience but additional considerations:

- **Conversation storage:** Interfaces store conversations in their databases for user convenience
- **Account security:** Security depends partly on account protection (strong passwords, multi-factor authentication)
- **User settings:** Some interfaces offer settings controlling data retention or opting out of certain uses
- **Visibility:** Interfaces make AI accessible to users who might not understand security implications

For individual professional use on less sensitive tasks, web interfaces with commercial accounts can be appropriate. For production systems processing confidential client data at scale, API access provides better control and security.

Reading the Fine Print

Not all commercial AI services are equal. When evaluating providers, examine:

1. **What exactly is prohibited:** Does the agreement prohibit all uses of your data, or just training? Are there exceptions for security monitoring, legal compliance, or service improvement?
2. **What happens to conversation history:** If using an interface that stores conversations, where is that data kept, how long, and who can access it?
3. **Deletion capabilities:** Can you delete data on demand? What happens when you delete conversations or close your account?
4. **Audit and logging:** What logs are maintained, for how long, and who can access them?
5. **Breach notification:** Does the agreement require notification if your data is breached?
6. **Liability provisions:** What recourse do you have if the provider breaches data handling terms?
7. **Third-party access:** Under what circumstances might the provider share your data with third parties or respond to the legal process?

Can LLMs Share Confidential Information?

This question troubles many legal professionals: even if commercial agreements prohibit data misuse, could the LLM itself somehow leak confidential information to other users? Could information you provide inadvertently appear in someone else's responses?

The technical answer is no, for reasons grounded in how LLMs actually work.

Why LLMs Can't Learn from Prompts

Recall from Chapters 1 and 2 that LLMs are static systems. Their neural network parameters are fixed after training completes. When you submit a prompt:

1. The LLM processes your input through its existing parameters
2. Generates a response based on those fixed parameters
3. Returns to exactly the same state afterward

Your prompt doesn't modify these parameters. The LLM cannot learn from your interaction. It has no mechanism to incorporate your information into its knowledge base.

This is fundamentally different from systems that learn continuously. Some AI systems do learn from user interactions, updating their models based on feedback. LLMs as currently deployed in commercial services don't work this way. They're inference-only systems—they apply their fixed knowledge to new inputs without modifying that knowledge.

Why Other Users Can't Access Your Data

The architecture prevents cross-user data exposure:

Isolated context windows: Each request gets its own context window. User A's context window never contains User B's data. The LLM processes requests independently.

No persistent memory: After processing your request, the LLM retains nothing. When the next user's request arrives (possibly to the same LLM instance), it starts from a clean state with only that user's context window visible.

Randomized routing: Requests are typically routed to available compute resources pseudo-randomly. Even if you submit multiple requests in sequence, they might go to different LLM instances. There's no persistent "session" at the LLM level.

No query mechanism: LLMs have no ability to search previous requests or access other users' data. They can only process what appears in the current context window.

Theoretical Vulnerabilities

While the architecture prevents inadvertent leakage, some theoretical vulnerabilities warrant mention:

Training data reconstruction: Academic research has shown that with extensive effort, attackers might extract small amounts of training data from LLMs by carefully probing them. However, this affects training data (public text from the internet), not user prompts, which aren't part of training in commercial services.

Prompt injection attacks: Sophisticated attackers might craft prompts that attempt to manipulate the system into revealing information about its operation or configuration. However, these attacks don't expose other users' data because that data isn't accessible to the LLM.

Interface vulnerabilities: Security issues in the interface layer (web application bugs, database vulnerabilities) could potentially expose conversation data stored by the interface. This is why interface security matters, though it's separate from LLM security per se.

Provider access: The AI provider's employees or systems might technically access your data if they breach their own policies or if their systems are compromised. This is a provider trustworthiness question rather than an LLM architecture question.

Comparison to Other Cloud Services

To contextualize the risk, consider that sending data to commercial LLM services presents security characteristics similar to other cloud services legal professionals routinely use:

Email services: When you email client information using Office 365, Gmail, or other cloud email, that data sits on provider servers indefinitely. Email systems maintain persistent storage of your communications. Staff at email providers could technically access your email if they violated policies. Yet legal



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

professionals use cloud email extensively, relying on contractual protections and provider trustworthiness.

- 1. Document management systems:** Cloud-based document management platforms store client documents on provider servers long-term. This presents a larger attack surface than LLM requests that exist in temporary memory for seconds.
- 2. Ediscovery platforms:** Discovery professionals routinely upload millions of potentially confidential documents to third-party ediscovery platforms. These platforms process, store, and allow access to documents for months or years during litigation.
- 3. Video conferencing:** Confidential client meetings occur over Zoom, Teams, or other platforms that route audio and video through provider infrastructure.
- 4. Cloud storage:** Documents stored in Dropbox, Box, Google Drive, or SharePoint reside on provider servers with provider employees potentially able to access them.

The legal profession has collectively determined that these cloud services, when used with appropriate commercial agreements and security measures, provide reasonable expectation of privacy sufficient to avoid confidentiality breaches or privilege waivers. The analysis for commercial AI services is similar.

Reasonable Expectation of Privacy

The legal standard typically governing whether using third-party services waives privilege or breaches confidentiality is “reasonable expectation of privacy.” This framework, developed through case law and bar association guidance, assesses whether the circumstances surrounding communication or data storage suggest the parties intended to maintain confidentiality.

Legal Framework

Courts have consistently held that using third-party technology services doesn’t waive privilege if users have reasonable expectation of privacy in their communications. This principle has been applied to:

Email: Using employer-provided or cloud-based email for privileged communications doesn’t waive privilege despite third-party hosting, as long as users reasonably expect communications to remain private.

Cloud storage: Storing privileged documents in cloud services doesn’t waive privilege when done through services providing appropriate security and confidentiality commitments.

Litigation support vendors: Sharing privileged documents with ediscovery vendors, forensic experts, or litigation support services doesn’t waive privilege because these disclosures are made in furtherance of legal representation with reasonable confidentiality expectations.

Factors Supporting Reasonable Expectation

Several factors support finding reasonable expectation of privacy when using commercial AI

services:

Contractual protections: Commercial agreements explicitly prohibiting data disclosure or use support reasonable privacy expectations. These contracts create enforceable obligations protecting confidentiality.

Industry-standard security: Using services that implement industry-standard encryption, access controls, and security measures indicates reasonable care to protect confidentiality.

Limited disclosure: Data shared with AI providers is disclosed only to the extent necessary for the service, not broadly published or made available to unrelated parties.

Purpose of disclosure: Using AI to analyze documents serves the purpose of legal representation, similar to engaging other technology providers or expert consultants.

Vendor reliability: Reputable AI providers with established track records, security certifications, and significant business incentives to protect customer data support reasonable privacy expectations.

ABA and State Bar Guidance

The ABA and various state bar associations have addressed technology use and confidentiality obligations:

ABA Formal Opinion 477R (Revised 2017): Addressed email and electronic communications, concluding that lawyers can generally use unencrypted email because they have reasonable expectation of privacy. The opinion emphasized that lawyers should understand the technology they use and take reasonable security measures.

ABA Formal Opinion 483 (2018): Discussed cloud storage, concluding that lawyers may use cloud storage with reasonable care to ensure confidentiality. Factors include understanding how providers will handle data, evaluating security measures, and staying informed about developments.

ABA Formal Opinion 512 (2024): Specifically addressed Generative AI, emphasizing lawyers' duties to understand tools they use and take reasonable measures to protect confidential information. The opinion didn't treat AI as categorically different from other cloud technologies regarding confidentiality.

State bar opinions: Various state bars have issued guidance generally consistent with ABA positions, emphasizing reasonable care, understanding of technology, and appropriate vendor selection rather than prohibiting cloud technology use.

Applying Standards to AI Services

Using commercial AI services through appropriate agreements satisfies reasonable expectation of privacy for several reasons:

More ephemeral than alternatives: Data sent to LLMs exists in temporary memory for sec-

onds, unlike email or documents stored indefinitely on cloud servers. The brief existence actually reduces exposure compared to persistent storage.

Strong contractual protections: Commercial AI agreements typically include specific prohibitions on data use, often stronger than generic cloud service terms.

No persistent storage in LLM: The LLM itself stores nothing. Only interface applications maintain conversation history, and this is typically deletable by users.

Security measures: Enterprise AI providers implement security measures comparable to or exceeding those of other legal technology providers.

Professional acceptance is growing: As legal professionals gain sophistication about AI security, using commercial AI services with appropriate safeguards is becoming accepted practice, much as cloud email and storage became accepted.

The key is using AI services properly—commercial agreements, understanding data handling, appropriate vendor selection, and reasonable security measures. Using consumer AI services without commercial protections for confidential client information would not satisfy reasonable expectations of privacy. Using commercial services with proper agreements likely does.

Special Considerations for Regulated Data

Certain types of data trigger additional regulatory requirements beyond general confidentiality obligations:

Healthcare Data (HIPAA)

Protected health information (PHI) under HIPAA requires specific safeguards. If your legal work involves healthcare data, ensure:

- AI vendors offer HIPAA-compliant configurations
- Business Associate Agreements (BAAs) are in place
- PHI is appropriately de-identified before processing, or
- Full HIPAA safeguards are implemented

Financial Data

Financial services data may be subject to regulations like GLBA (Gramm-Leach-Bliley Act) or industry-specific requirements. Financial institutions using AI services should verify compliance with applicable regulations.

International Data

Data from European individuals triggers GDPR requirements. Data from other jurisdictions may trigger additional data protection laws. Consider:

- Where data is processed and stored

- Data transfer mechanisms for cross-border data movement
- Individual rights under applicable privacy laws
- Documentation and consent requirements

Government Classified Information

Classified government information requires specific security clearances and approved systems. Standard commercial AI services are not appropriate for classified data. Government agencies must use approved systems meeting classification-level requirements.

Trade Secrets

While trade secrets don't trigger specific regulations like healthcare data, their value demands careful protection. Consider whether the benefits of AI analysis outweigh risks of exposure, even through secure commercial services.

Looking Ahead

This chapter examined security, privacy, and confidentiality considerations when using AI services. We explored how AI systems handle data, what protections commercial agreements provide, why the technical architecture prevents inadvertent disclosure, and how using AI services compares to other cloud technologies legal professionals routinely employ.

The conclusion many discovery professionals have reached is that commercial AI services, used appropriately, present manageable security risks comparable to other cloud technologies. The key lies in understanding the technology, selecting reputable providers with strong contractual protections, implementing reasonable security measures, and training users about appropriate use.

Security and confidentiality concerns, while legitimate, need not prevent AI adoption. They should inform how you adopt AI—driving thoughtful vendor selection, appropriate safeguards, and reasonable security practices.

With concerns about data security addressed, we can now turn to the broader ethical framework governing AI use in legal practice. Chapter 5 examines the comprehensive ethical obligations lawyers face when using AI, synthesizing guidance from the ABA and state bar associations into practical frameworks for responsible AI deployment. Security is one component of ethical AI use, but the ethical considerations extend further, encompassing competence, supervision, transparency, billing, and professional judgment.

05

Ethical and Professional Responsibilities

CHAPTER FIVE

Ethical and Professional Responsibilities

The integration of Generative AI into legal practice presents ethical questions that extend far beyond security and accuracy. These questions touch the core of what it means to be a competent legal professional in an era of rapidly advancing technology. Can lawyers delegate legal analysis to AI systems? What must they disclose to clients about AI use? How should they bill for AI-assisted work? What supervisory obligations arise when staff use these tools?

These questions aren't hypothetical exercises. They affect daily practice decisions, client relationships, professional reputations, and potentially bar licenses. The legal profession has begun developing frameworks to address them, though the guidance continues to evolve as both the technology and our understanding of its implications mature.

This chapter synthesizes ethical guidance from the ABA and state bar associations into practical frameworks for responsible AI use. We examine core obligations, explore how they apply to specific AI applications in discovery, and provide decision-making tools for navigating ethical uncertainties. The goal is not to provide definitive answers to every situation—the technology evolves too rapidly for that. Instead, we aim to establish principles and analytical approaches that remain relevant as capabilities advance.

Why Ethics Matter Differently in Legal AI

The legal profession operates under distinctive constraints that shape how we must approach technology adoption. Lawyers serve as fiduciaries to their clients, owing duties of loyalty, confidentiality, and competent representation that exceed ordinary commercial relationships. These heightened obligations mean lawyers cannot simply adopt useful technology the way other professionals might. Each technology decision must be evaluated through the lens of client interests and professional responsibilities.

Legal work often occurs in adversarial contexts where opposing parties scrutinize work product for weaknesses. If you use AI for document review, opposing counsel might challenge your methodology. If AI assists in brief writing, courts might question accuracy and attribution. These challenges don't mean AI use is inappropriate, but they demand defensible implementation and clear documentation.

ABA Formal Opinion 512: Comprehensive Analysis

On July 29, 2024, the ABA Standing Committee on Ethics and Professional Responsibility issued Formal Opinion 512, providing the profession's first comprehensive ethical guidance specifically addressing Generative AI use. This landmark opinion doesn't create new ethical rules but interprets existing Model Rules of Professional Conduct as they apply to AI technology.

Understanding Opinion 512 is essential for any legal professional using or considering AI tools. The opinion addresses six core areas: competence, confidentiality, verification, transparency,

supervision, and billing. We examine each in detail.

1. Duty of Competence

The Rule: Model Rule 1.1 requires lawyers to provide competent representation, defined as “the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation.”

Application to AI: The ABA interprets this duty to require that lawyers using AI tools develop a reasonable understanding of how those tools work, what they can and cannot do reliably, and what risks they present. Notably, the opinion emphasizes that lawyers need not become AI experts or understand technical implementation details. What’s required is functional understanding sufficient to use tools effectively and recognize their limitations.

Practical Implications:

Lawyers using AI should understand what the specific tool does, what types of outputs it generates, known limitations and failure modes, appropriate use cases, and how to verify outputs for accuracy. The competence duty extends to staying informed about technological developments through CLE programs, professional publications, consulting with technology experts when implementing new tools, and testing tools on representative tasks before production deployment.

Competence requirements vary by practice area. Discovery professionals need deeper understanding of AI document analysis capabilities than general practitioners. Competent lawyers recognize when tasks exceed their technological understanding and seek assistance—consulting technology specialists, working with vendors, or declining to use AI for applications beyond current comprehension.

2. Confidentiality and Data Protection

The Rule: Model Rule 1.6 requires lawyers to maintain client confidentiality, prohibiting disclosure of information related to representation without client consent except in limited circumstances.

Application to AI: Opinion 512 concludes that lawyers may use AI tools that process confidential information, but must take reasonable steps to protect that information from unauthorized disclosure.

Practical Implications:

Before sending confidential information to AI services, lawyers must evaluate data handling practices. This evaluation should address what the vendor does with customer data, whether data is used for training or improvement, how long data is retained, what security measures protect data, whether the vendor complies with relevant regulations, and what happens if the vendor experiences a breach.

Using consumer AI services for confidential information likely violates Rule 1.6 because these

services may use data for training. Commercial services with appropriate agreements generally don't raise the same concerns. Lawyers must implement reasonable security measures when using AI, which might include using services with strong encryption, implementing access controls, enabling multi-factor authentication, using private deployments for highly sensitive matters, and training staff on security practices.

What's "reasonable" depends on sensitivity of information, nature of the tool, and available security options. Highly sensitive matters (national security, trade secrets, high-stakes litigation) might warrant more stringent measures than routine matters.

3. Independent Verification and Due Diligence

The Rule: Model Rule 1.1's competence requirement implies that lawyers must ensure work product meets professional standards regardless of what tools assisted in creation.

Application to AI: Opinion 512 emphasizes that lawyers cannot blindly rely on AI outputs. They must independently verify AI-generated content before using it in representation. The opinion specifically references cases where lawyers were sanctioned for citing fake cases generated by ChatGPT, using these incidents to illustrate verification imperatives.

Practical Implications:

The degree of verification needed depends on how AI outputs are used and what risks errors present. Court filings require thorough verification—case citations must be checked, factual claims must be verified against sources, legal arguments must be evaluated for soundness. Client communications should be verified to the degree appropriate for the subject matter. Internal work product might require less immediate verification, but anything used for decision-making must be verified before reliance. Document review AI classifications require validation through sampling and quality control processes.

Verification should address factual accuracy, citations (verify that cited authorities actually exist and say what's claimed), logical soundness (evaluate whether reasoning is sound), completeness (consider whether AI missed important issues), and context appropriateness (assess whether outputs properly account for case-specific circumstances).

Verification methods include source checking for any citations or references, cross-referencing AI outputs against other sources, having subject matter experts review AI-generated content in their specialties, sampling large volumes of AI outputs to assess reliability, and considering how opposing counsel might challenge AI-assisted work. Document verification efforts, particularly for work that might be scrutinized, as notes provide evidence of due diligence.

4. Transparency and Client Communication

The Rule: Model Rules 1.4 (Communication) and 8.4 (Misconduct) create obligations to communicate with clients and to be truthful in professional dealings.

Application to AI: Opinion 512 suggests lawyers should consider disclosing AI use to clients, particularly when it materially affects representation or clients might reasonably expect to know. The opinion stops short of mandating disclosure in all circumstances, recognizing that

routine tool use might not require disclosure any more than using word processors or email requires specific client notification.

Practical Implications:

Disclosure is advisable when using novel or experimental tools not yet standard in practice, when AI substantially influences analysis or strategy, when clients have voiced concerns about AI or technology, for highly sensitive matters, or when AI use creates risks clients might not anticipate. Some lawyers include AI use disclosure in engagement letters or matter-specific communications.

Communication should explain in accessible terms what AI tools do and how they'll be used, describe benefits like efficiency or cost savings, acknowledge limitations and safeguards that address them, explain how confidential information will be protected, and clarify that humans remain responsible and will verify outputs.

Disclosure may not be necessary when AI use is comparable to other standard tools (like legal research databases) or plays a minor role in work product. As AI use becomes standard practice, disclosure requirements likely diminish.

Regarding disclosure to courts and opposing counsel, there's no general disclosure duty to announce AI use. However, the duty of candor to tribunals requires that statements to courts be accurate regardless of what tools produced them. In discovery, if opposing counsel asks about methodologies or tools used, truthful responses are required. Some courts have adopted local rules requiring disclosure of AI use in filings, and lawyers must comply with applicable rules.

5. Supervisory Responsibilities

The Rule: Model Rules 5.1 (supervision of lawyers) and 5.3 (supervision of nonlawyer assistants) require lawyers with managerial authority to ensure that subordinates' conduct complies with professional obligations.

Application to AI: Opinion 512 extends supervisory obligations to AI tool use. Partners and legal managers must ensure that lawyers and staff using AI do so competently, ethically, and consistently with the firm's obligations to clients.

Practical Implications:

Law firms and legal departments should develop clear policies governing AI use. These policies might address approved tools and what AI services may be used under what circumstances, prohibited uses (e.g., consumer AI services for confidential matters), vendor approval processes, data handling requirements, verification standards, documentation requirements, and escalation procedures.

Supervisory lawyers must ensure subordinates receive adequate training on ethical obligations, competent use of specific tools, security practices for protecting confidential information, verification techniques, and how to report concerns or errors.

Supervision isn't one-time training but ongoing oversight through quality review (sampling

work product to ensure AI is used appropriately), compliance monitoring (checking that policies are followed), staying informed about changes to tools in use, addressing issues when they arise, and refining policies based on experience.

6. Billing Practices

The Rule: Model Rules 1.5 (Fees) and 8.4 (Misconduct) prohibit charging unreasonable fees or billing fraudulently.

Application to AI: Opinion 512 addresses how AI use affects billing, particularly for hourly billing arrangements. The opinion concludes that lawyers cannot bill clients for time spent learning to use AI tools, and that efficiency gains from AI should generally benefit clients through reduced fees rather than increased profits through maintained billing at pre-AI rates for faster work.

Practical Implications:

Lawyers cannot bill for time spent learning to use AI tools (analogous to not billing clients for learning law), troubleshooting AI tool problems that didn't arise from case-specific circumstances, or experimenting with new AI approaches without authorization.

If AI allows completing tasks faster, billing should generally reflect actual time invested, not pre-AI time estimates. For example, if document review that previously took 40 hours now takes 10 hours with AI assistance, billing 40 hours would be inappropriate. For alternative fee arrangements (fixed fees, success fees), AI efficiency gains might benefit the firm rather than requiring fee reductions, depending on the agreement.

Consider addressing AI use in fee agreements, particularly for hourly billing. This provides transparency about how fees are calculated. Clients expecting traditional hourly billing may need education about how AI affects time investment and billing.

Additional Ethical Considerations

Bias and Fairness

LLMs learn from training data that reflects biases present in society and historical texts. This creates risks of perpetuating or amplifying bias in legal work. Training data may overrepresent certain demographics or perspectives, legal texts reflect historical practices that may have been discriminatory, and language patterns can encode subtle biases about race, gender, socioeconomic status, and other characteristics.

In legal AI applications, bias might manifest as biased document classification that systematically miscategorizes documents related to certain groups, predictive analysis that reflects historical biases in judicial decisions, AI-drafted content that inadvertently uses biased language or framing, or pattern recognition that flags behavior more common in certain demographic groups.

Lawyers using AI should understand potential bias risks in their specific applications. Before deploying AI at scale, test for biased outcomes across demographic categories. Ongoing mon-

itoring can detect bias that wasn't apparent in initial testing. When bias is detected, take steps to mitigate it by adjusting prompts, using different tools, or increasing human oversight. Document bias analysis and mitigation efforts, particularly for applications in sensitive contexts. Criminal justice applications deserve special scrutiny. AI used in matters affecting liberty, employment discrimination cases, housing discrimination, or other civil rights contexts requires heightened attention to bias risks.

Professional Judgment and Independence

Model Rule 2.1 requires lawyers to exercise independent professional judgment and render candid advice. AI tools, no matter how sophisticated, cannot replace this judgment.

Treat AI outputs as information to inform judgment, not as final answers. Approach AI recommendations with the same critical thinking applied to any other source. Consider case-specific factors AI might not fully account for. Ensure decisions prioritize client interests, not merely what AI suggests is efficient or likely. Don't delegate decision-making to AI even when its suggestions seem reasonable. Vendor marketing shouldn't drive adoption of AI tools not actually beneficial for clients. Pressure to use AI for efficiency can't override professional judgment about when AI is inappropriate.

Certain decisions particularly require human judgment: settlement recommendations, litigation strategy choices, risk assessments for clients, advice on transaction structures, and ethical decisions within representation. AI can inform these decisions but cannot make them. The lawyer's judgment, informed by experience, knowledge of the client, understanding of context, and professional responsibility, must remain central.

State Bar Guidance

While Opinion 512 provides the ABA's comprehensive framework, individual state bars govern attorney conduct within their jurisdictions. Several states have issued their own guidance, including California, Florida, New York, and the District of Columbia. The guidance from these states largely aligns with Opinion 512, emphasizing that existing ethical rules apply to AI use and require competence, confidentiality protection, verification, appropriate supervision, and reasonable billing.

States sometimes emphasize issues particularly relevant to their jurisdictions—privacy laws, specific court rules, or local practice norms—but the core principles remain consistent: AI is permitted but must be used competently, ethically, and with appropriate safeguards. Practitioners should check their own jurisdiction's guidance for any specific requirements or interpretations that apply to their practice.

Looking Ahead

This chapter examined the ethical framework governing AI use in legal practice, synthesizing guidance from the ABA and state bar associations into practical decision-making tools. The core principles—competence, confidentiality, verification, transparency, supervision, and reasonable billing—provide foundation for responsible AI adoption. These principles prepare us for Part Two, where we explore advanced AI capabilities including agentic systems that can plan multi-step workflows, use tools strategically, and iterate toward goals autonomously.



PART TWO

AGENTIC AI AND REASONING MODELS

PART TWO

Agentic AI and Reasoning Models

Part One equipped you with foundational knowledge about Large Language Models—what they are, how they work, their limitations, and the ethical framework governing their use. Now we turn to two major evolutions in legal AI that move beyond simple prompt-response interactions: agentic systems that can plan and execute complex multi-step workflows autonomously, and reasoning models that can think through difficult problems methodically before responding.

These advances represent a fundamental shift in how AI assists legal work. Agentic systems don't just respond to questions—they pursue goals, breaking assignments into steps, executing searches, analyzing results, identifying gaps, and iterating until they produce comprehensive work products. Reasoning models don't just generate immediate responses—they engage in extended deliberation, working through logical chains, considering multiple perspectives, and evaluating their own conclusions before answering.

For discovery professionals, these capabilities address critical needs. Real discovery investigations don't follow linear paths—you search for information about Topic A, which reveals Custodian B is important, leading you to their documents, which mention Event C you hadn't known about. Human investigators naturally follow these threads. Simple LLM queries can't, but agentic systems can. Similarly, privilege determinations and credibility evaluations require working through multiple considerations systematically. Quick responses often miss important nuances. Reasoning models provide the deliberate analysis these tasks require.

What You'll Learn

Part Two explores how agentic systems use goal-directed behavior, multi-step planning, tool use, and iterative refinement to conduct investigations rather than just answer questions. You'll understand how reasoning models employ extended chain-of-thought deliberation to work through complex legal analysis, contradictory evidence, and strategic decisions. We examine when to deploy each capability, how to maintain quality control and oversight, and what ethical obligations intensify as AI systems become more autonomous and sophisticated.

Why These Capabilities Matter

Discovery work inherently involves both complex multi-step processes and difficult analytical problems. Many discovery questions can't be answered by reading single documents—they require synthesizing information across dozens or hundreds of sources, identifying patterns, recognizing contradictions, and building coherent narratives from fragments. Agents can orchestrate this synthesis in ways that simple prompts cannot. And as investigations progress, understanding deepens and priorities shift. Agentic systems can adapt to evolving contexts, refining their approaches based on what they've learned, while reasoning models can reconsider earlier conclusions in light of new information.

These advanced capabilities promise dramatic transformation. Tasks requiring days of human analysis might be accomplished in hours. Investigations demanding extensive manual synthesis could be conducted systematically at scale. But these capabilities also create challenges.

Autonomous systems operating with less granular human oversight require different quality assurance approaches. Verifying that an agent properly conducted a multi-step investigation or that a reasoning model correctly evaluated a complex legal question proves more complex than verifying a simple document summary. The ethical questions about professional judgment and human oversight become more acute when systems operate autonomously rather than simply responding to prompts.

How to Approach Part Two

Agentic AI and reasoning models represent the cutting edge of current legal technology. Not every discovery professional needs to implement these systems immediately. But understanding these capabilities prepares you to evaluate vendor claims, plan technology roadmaps, and stay current as the technology evolves. The trajectory of legal AI development points clearly toward both increased autonomy and more sophisticated reasoning.

As you read Part Two, continue thinking about your practice and workflows. Where do you encounter complex multi-step tasks that could benefit from autonomous AI assistance? What analytical problems require careful reasoning through multiple considerations? What investigations require following threads through multiple sources? These are the applications where advanced AI capabilities shine—and where Part Two’s insights will prove most valuable.

Looking Ahead

Part Two explores territory that many legal professionals haven’t yet encountered directly. These advanced AI capabilities remain newer and less widely deployed than simpler LLM applications. But they represent where legal AI is moving, and understanding them positions you to lead rather than follow as capabilities mature and become more accessible. The technical foundation established here prepares you for the practical applications demonstrated in Part Three, where we show these advanced systems transforming actual discovery workflows.

Let’s begin by examining what makes AI systems “agentic” and how autonomous assistance differs from simple prompt-response interactions.

06

Understanding Agentic AI Systems

—

CHAPTER SIX

Understanding Agentic AI Systems

In Chapter 1, we described Large Language Models as powerful prediction engines that generate text by anticipating what word should come next in a sequence. In Chapters 2 through 5, we explored how these systems work, their limitations, and the frameworks for using them responsibly. Throughout Part One, we examined LLMs as responsive systems—you provide a prompt, the LLM generates a response, and the interaction concludes.

Agentic AI systems operate fundamentally differently. Rather than simply responding to prompts, they pursue goals through autonomous planning and execution. Instead of a single request-response cycle, agentic systems orchestrate dozens or hundreds of LLM calls, each building on previous results, adjusting strategies when approaches fail, and iterating toward objectives without constant human direction.

This chapter explores what makes AI systems “agentic,” examining the components that enable autonomous behavior, the architectural patterns that organize these components, and the practical implications for discovery professionals considering these advanced capabilities. Chapter 7 then examines reasoning models—a complementary advance that enables extended deliberation on complex analytical problems. Together, these capabilities represent the next evolution in AI-assisted discovery work.

Defining Agentic AI

The term “agentic” derives from “agency”—the capacity to act independently and make autonomous decisions toward goals. Agentic systems differ from traditional LLM applications in several key ways.

Goal-oriented operation and planning: Traditional LLM interactions are task-oriented. You specify a task (“Summarize this document”), and the system completes it. Agentic systems are goal-oriented. You specify an objective (“Identify all relevant evidence”), and the system determines what tasks are needed to achieve that objective.

This distinction matters because goal achievement requires planning and adaptation, not just following instructions. When you assign an agent a goal, it first analyzes what’s needed, breaks the objective into manageable subtasks, and sequences those tasks logically.

Tool use capability: While LLMs generate text, agentic systems can invoke external tools—search functions, data extraction utilities, calculation capabilities, file operations, or specialized analysis tools. Critically, agents decide which tools to use and when, based on task requirements.

Self-improvement through iteration: Agentic systems evaluate their own outputs and refine them through multiple passes. After completing initial analysis, an agent might identify gaps in its work, recognize that conclusions lack sufficient support, or determine that alternative

approaches might yield better results. The agent can then conduct additional searches, re-analyze information, or approach problems differently.

Related to this iterative capability is error detection. When agents make mistakes—searching with terms that return irrelevant results or drawing incorrect inferences—they can often detect these errors and take corrective action.

State management: While individual LLMs are stateless (Chapter 2), agentic frameworks maintain state across multiple operations. The agent “remembers” what it’s already done, what it learned from previous steps, and how current work connects to the broader goal. This allows agents to conduct coherent investigations despite being built on stateless LLMs.

What Agentic Systems Are Not

Understanding agentic AI also requires clarity about what it isn’t:

Not autonomous AGI: Agentic AI systems are not artificial general intelligence (AGI) capable of human-like reasoning across all domains. They’re specialized tools designed for specific types of tasks, operating within defined boundaries.

Not learning systems: Current agentic systems don’t learn from experience in the way humans do. Each investigation starts fresh. The agent doesn’t remember previous matters or accumulate expertise over time. (Some architectures maintain learnings within single sessions, but this differs from persistent learning across sessions.)

Not independent decision-makers: Despite autonomy within workflows, agents operate under human direction. Humans set goals, approve deployment, and remain responsible for outcomes. Agents are sophisticated tools, not independent practitioners.

Not replacements for human judgment: Agents can execute complex investigations, but professional judgment about strategy, significance, and next steps remains human responsibility. We explore this extensively in Chapter 9’s ethical discussion.

Not infallible: Agentic systems make mistakes. They can pursue unproductive directions, miss important information, or draw incorrect conclusions. Autonomy doesn’t equal accuracy. All the limitations from Part One—hallucinations, context window constraints, training cutoffs—still apply.

What’s Happening Under the Hood

To understand how agentic systems actually work, consider what happens when you assign a discovery goal: “Identify all documents related to the patent dispute with TechCorp.”

A traditional search would return documents matching “patent dispute TechCorp” and stop. An agentic system pursues the goal through multiple rounds of thinking, acting, and observing:

Round 1: Think: “I need to start with broad patent-related terms.” Act: Search for “patent TechCorp.” Observe: Results mention Patent #7,234,567, cite three key custodians (Sarah Chen, Michael Rodriguez, James Park), and reference a March 2023 timeframe.

Round 2: Think: “Now I should search specifically for that patent number and those custodians.” Act: Search for “Patent 7234567” combined with each custodian name. Observe: These emails frequently mention “Project Aurora” and a meeting on March 15, 2023. Several reference “prior art concerns.”

Round 3: Think: “Project Aurora appears to be the internal code name. I should search for that term around the March 15 meeting date.” Act: Search for “Project Aurora” in documents dated February-April 2023. Observe: Calendar invites for March 15 meeting include two external counsel and a patent examiner. Meeting notes indicate discussion of potential infringement.

Round 4: Think: “The external counsel names suggest outside firms were involved. I should identify their communications.” Act: Search for communications from those law firms. Observe: Opinion letters from counsel dated April 2023 provide formal patent analysis.

This process continues—potentially through dozens of rounds—until the system determines it has comprehensively addressed the goal. Each observation informs the next action. The system follows leads autonomously, adjusts strategy based on what it finds, and builds understanding iteratively rather than delivering a single set of search results.

The key insight: You’re not directing each step. You set the goal, and the system determines how to achieve it through systematic investigation.

It Takes Time

Understanding the time dimension of agentic systems is crucial. Unlike traditional search that returns results in seconds, an agentic system might work for several hours—or even days—pursuing an investigation goal. During this time, it may conduct hundreds of searches, analyze thousands of documents, follow dozens of leads, and build comprehensive understanding before presenting findings.

This extended timeframe enables thoroughness impossible with traditional tools. The system can pursue every lead, verify conclusions across multiple sources, identify gaps in its analysis, and conduct additional searches to address those gaps. It can build a complete picture rather than providing preliminary results that require human investigation to complete.

For discovery professionals, this means rethinking workflow. Rather than conducting multiple rounds of manual searches over days or weeks—each requiring analysis before determining the next search—you can assign comprehensive investigation goals and receive analyzed intelligence. The trade-off is patience: you must allow the system time to work rather than expecting immediate results.

Why These Capabilities Matter for Discovery

Understanding what agentic systems are—and are not—prepares us to examine why these capabilities matter for discovery work. Real investigations don’t follow linear paths. You search for Topic A, which reveals Custodian B, whose documents mention Event C. Each discovery shapes the next inquiry. Simple LLM interactions can’t pursue these emergent leads autonomously. Agentic systems can.

Many discovery questions require synthesizing information scattered across dozens of documents. “What did the company know about the defect, and when?” can’t be answered by any single email, report, or memorandum. Each source provides one piece; the answer emerges through systematic analysis across the entire set. Agentic systems orchestrate this process autonomously—following threads, identifying patterns, recognizing contradictions, and building coherent understanding.

Consider the goal “Identify all documents related to the patent dispute.” A simple LLM query might search for “patent dispute” and return results. An agentic system might plan:

- Search for documents using relevant patent terms
- Analyze results to identify specific patents, parties, and time periods
- Conduct targeted searches for the specific patents and custodians identified
- Build a chronology of key events and identify information gaps
- Conduct additional searches to address those gaps
- Synthesize findings into a comprehensive report

The agent creates this plan autonomously, adjusting as execution reveals new information or approaches prove ineffective.

These systems can process massive document volumes while maintaining sophisticated analysis and incorporating quality control. Rather than producing search results for human review, they conduct investigations and deliver analyzed intelligence. This doesn’t replace professional judgment about strategy and significance, but it expands what’s possible in terms of comprehensiveness, speed, and consistency.

Components of Agentic Systems

Agentic AI systems comprise several technical components working together. Understanding these components at a high level helps you evaluate agentic tools and recognize their capabilities and limitations. An orchestration layer coordinates planning, execution, memory, and evaluation components to enable goal-directed behavior.

Planning: The planning component receives goals and creates action plans by prompting an LLM to analyze the goal and generate steps. The system considers available tools, breaks the goal into manageable tasks, and sequences those tasks logically.

Different systems use different planning approaches. Some generate complete plans upfront and execute all steps in sequence—effective for well-defined problems. Others create high-level plans initially and fill in details as work proceeds, adapting to discoveries. Still others generate each next step based on current circumstances without extensive forward planning. For discovery applications, approaches that combine strategic planning with flexibility to adapt work best.

Execution and decision-making: The execution component carries out plan steps by invoking tools or prompting LLMs to perform specific tasks. When a step requires using a tool—searching documents, extracting data, or performing calculations—the system identifies the appropriate tool, formats the input, invokes the tool, and processes the output.

During execution, agents must make numerous tactical decisions: Which tool to use? What

search terms are likely effective? Do results answer the current question? Should the plan be modified based on new information? The system makes these decisions by prompting LLMs with the decision context and criteria, leveraging their reasoning capabilities to select appropriate next steps.

After each action, the execution component updates the system's state—marking completed steps, recording outputs, updating understanding based on new information, and determining what comes next. This state management allows the agent to maintain coherent progress despite being built on stateless LLMs.

Memory: Memory components maintain information across the multiple LLM calls that comprise an agentic workflow. This working memory holds the goal, the current plan, what's been accomplished, results from previous steps, and items requiring further investigation.

Since context windows are limited (Chapter 2), memory systems must manage what information gets included in prompts. They summarize extensive previous work to fit available space, prioritize the most relevant information for the current task, and retrieve specific details when needed. This allows agentic systems to conduct extensive investigations despite context window constraints.

Evaluation and verification: As we emphasized earlier, verification is critical for any AI outputs. Agentic systems can incorporate verification into their workflows by prompting LLMs to critically assess work product. Agents can evaluate whether search results actually relate to search terms, whether conclusions follow from evidence, whether there are gaps or inconsistencies, and whether outputs address the original goal.

More sophisticated verification involves cross-checking across sources—confirming that multiple independent sources support conclusions, identifying contradictions, verifying that citations reference actual documents, and ensuring extracted data remains consistent across mentions. These verification mechanisms help identify outputs requiring additional human review.

These components work together under an orchestration layer that manages overall workflow—initiating planning, sequencing execution, handling errors, determining when goals are achieved, and compiling final outputs. When problems arise, the orchestration layer determines appropriate responses: retry the operation, try an alternative approach, or escalate to human oversight. This coordination enables agentic systems to pursue complex investigations while adapting to challenges and maintaining progress toward goals.

Agentic Architectures

These components can be organized in various architectural patterns, each with different strengths and appropriate use cases.

Single-Agent Architecture

The simplest agentic architecture involves one agent with all components integrated. A single agent receives goals, plans approaches, executes steps, maintains state, and produces outputs. This design offers important advantages: it's simpler to build and maintain, maintains

consistency through a single perspective, operates efficiently without coordination overhead, and provides transparency since you're tracking one agent's reasoning rather than multiple interacting agents.

However, single agents face inherent limitations. They can't be simultaneously optimized for diverse tasks, everything runs sequentially rather than in parallel, and very complex goals may overwhelm single-agent planning capabilities. For discovery applications, single-agent systems work well for document analysis within coherent scope, timeline construction from clear sources, targeted investigations with defined boundaries, and tasks requiring consistent analytical perspective. A single agent investigating communications between parties can plan search strategies, execute searches, analyze results for relevance, identify key items for detailed review, extract critical information, synthesize findings, identify gaps, conduct supplemental searches, and compile a comprehensive report—all while maintaining context and coherence throughout.

Multi-Agent Architecture

Multi-agent systems employ specialized agents, each optimized for particular tasks, working together on complex goals. Rather than one agent handling everything, multiple agents collaborate—each focusing on what it does best. For discovery applications, this specialization proves valuable. A search agent might excel at finding relevant documents across large collections, while an analysis agent focuses on extracting key information. A synthesis agent then combines findings into coherent narratives, while a quality assurance agent verifies accuracy and completeness.

The key benefit is that each agent can be optimized for its specific role, leading to higher quality results than a single generalist agent could achieve. However, coordinating multiple agents requires sophisticated orchestration to ensure coherent output and manage workflow complexity.

Hierarchical Agent Architecture

Hierarchical systems organize agents in tiers, with high-level strategic agents directing lower-level tactical agents. Strategic agents determine overall approach and break goals into major sub-goals. Tactical agents plan specific approaches to achieve those sub-goals. Execution agents perform concrete tasks like searching documents or extracting data.

This hierarchical approach scales well to very complex investigations—such as enterprise-scale discovery spanning multiple matters, or complex investigations requiring both strategic oversight and detailed execution across diverse document types and time periods. The structure allows for parallel processing while maintaining strategic coherence.

Control Mechanisms and Guardrails

Autonomous systems require controls ensuring they pursue goals appropriately without running indefinitely or producing unacceptable outputs. These guardrails protect against runaway execution while maintaining the benefits of autonomous operation.

Key control mechanisms include iteration limits to prevent endless refinement, scope con-

straints to keep investigations focused on relevant materials, and quality thresholds ensuring outputs meet professional standards. Human checkpoints provide strategic oversight at key milestones, while escalation triggers automatically involve human oversight when agents encounter uncertainty or contradictions.

These controls ensure that agentic systems remain powerful tools under human direction rather than independent decision-makers. The goal is to harness autonomous capabilities while maintaining professional oversight and responsibility for outcomes.

Knowing When Agents Are Overkill

Agentic approaches represent the most sophisticated AI capabilities currently available for discovery. That sophistication brings costs: longer processing time, higher computational expense, and more complex verification requirements. For many discovery tasks, simpler tools remain more appropriate.

Use traditional tools when:

- The task is straightforward and well-defined: “Summarize these 10 depositions” or “Find all emails mentioning Product X.”
- You need immediate results rather than comprehensive investigation.
- The document set is small enough for manual review or simple search.
- The question requires a single search or analysis pass rather than iterative investigation.

Consider agentic approaches when:

- The investigation requires following leads across multiple sources: “What did the company know about the safety issue, and when did they know it?”
- You need synthesis across large document sets: “Build a complete timeline of the acquisition negotiations.”
- The task involves identifying patterns or connections not apparent from individual documents.
- Comprehensive analysis would require dozens of search iterations if done manually.

The goal isn’t to use the most advanced AI available—it’s to match the tool to the task. A summary agent working for hours on a 5-document set makes no sense. That same agent pursuing a complex investigation across 50,000 documents might be exactly what’s needed.

As these tools become more common, part of discovery competence will be recognizing which approach serves each situation. Sometimes the right answer is still basic keyword search followed by human review. Understanding agentic capabilities means knowing when to use them and when simpler methods suffice.

Looking Ahead

This chapter introduced the core concepts of agentic AI—what makes systems “agentic,” what components enable autonomous behavior, how those components can be architecturally organized, and what control mechanisms ensure appropriate operation.



Understanding these foundations prepares you for Chapter 7, where we examine reasoning models and their role in complex analytical problems. While agentic systems excel at multi-step workflows and autonomous investigation, reasoning models provide extended deliberation on difficult questions requiring systematic evaluation. Following that technical foundation, Part Three demonstrates both capabilities through detailed discovery applications—showing how the architectures and components we’ve discussed here transform actual legal investigations.

Chapter 8 will demonstrate that agentic capabilities aren’t merely theoretical possibilities—they’re practical tools already transforming how discovery professionals approach complex investigations. The architectures and components we’ve discussed here provide the foundation for workflows that dramatically expand what’s possible in document analysis, investigation, and synthesis at scale.

07

Reasoning Models and Deep Analysis

—

CHAPTER SEVEN

Reasoning Models and Deep Analysis

In September 2024, a new category of AI model emerged that changed how we think about machine intelligence. OpenAI released the beta of its o1 model (“strawberry”), followed by similar capabilities from other providers, introducing what we now call reasoning models. These systems represent a fundamental shift in how AI approaches complex problems.

Unlike standard large language models that generate responses immediately, reasoning models take time to think through problems step by step. They work through complex logical chains, consider multiple perspectives, and evaluate their own conclusions before responding. For discovery professionals, this capability opens new possibilities for analyzing intricate legal questions, evaluating conflicting evidence, and conducting thorough investigations.

What Are Reasoning Models?

Standard LLMs produce responses in a single forward pass. You ask a question, and the model generates an answer token by token, completing the response in seconds. This approach works well for many tasks, but it has limitations when dealing with complex analytical problems that benefit from deliberate, step-by-step reasoning.

Reasoning models use a different approach. When you submit a query, these models engage in extended internal thinking before responding. They break problems into components, explore different analytical paths, identify potential errors in their reasoning, and refine their approach. This process, often called chain-of-thought reasoning, mirrors how human experts work through difficult problems.

The key distinction is deliberation. Standard models are fast but don’t reconsider their initial analytical path. Reasoning models take more time but produce more thorough, logically consistent analysis. For a simple question like summarizing a single document, this extra processing offers little benefit. For complex tasks like analyzing contradictory testimony across dozens of depositions or evaluating multi-factor privilege determinations, the improvement can be substantial.

The Thinking Process

When you use a reasoning model, the system goes through several phases:

Problem decomposition: The model breaks your question into logical components. For a discovery task like “identify all documents that support the plaintiff’s theory of conspiracy,” it might consider what evidence would demonstrate conspiratorial conduct, what time period matters, which custodians would be involved, and what communication patterns would be relevant.

Analytical exploration: The model considers multiple approaches to the problem. It might

evaluate different legal frameworks, consider various interpretations of ambiguous facts, or explore alternative explanations for patterns in the evidence.

Self-correction: As it works through the problem, the model identifies potential errors or gaps in its reasoning. It might recognize that an initial conclusion conflicts with other evidence, or that it has overlooked an important consideration.

Synthesis: The model integrates its various analytical threads into a coherent response, weighing different factors and qualifying conclusions where appropriate.

This internal process can take from a few seconds to several minutes, depending on the complexity of the task. Some systems show you a summary of this thinking process, while others keep it hidden. Either way, the result is typically more thorough and logically consistent than what standard models produce.

When Reasoning Helps in Discovery

Not every discovery task benefits from extended reasoning. Simple document summaries, straightforward searches, and routine categorization work fine with standard models. Reasoning models excel at tasks that require:

Multi-step logical analysis: Problems that require working through several layers of reasoning benefit from the model's ability to maintain logical consistency across multiple steps. Evaluating whether a document is privileged, for instance, might require analyzing the nature of the communication, the relationship between parties, the subject matter, and whether any exceptions apply.

Contradictory evidence: When testimony conflicts or documents present inconsistent information, reasoning models can methodically work through the contradictions, identify which sources are most reliable, and explain the discrepancies.

Complex pattern recognition: Identifying subtle patterns across large document sets, such as unusual communication sequences that might indicate coordination or concealment, requires the kind of deliberate analysis that reasoning models provide.

Legal and factual synthesis: Tasks requiring integration of legal standards with complex factual scenarios benefit from reasoning models' ability to work through multiple considerations systematically.

Practical Applications in Investigation and Discovery

Let's examine how reasoning models enhance several common discovery tasks:

Privilege Analysis

Privilege review represents one of the most complex analytical tasks in discovery. A document might be privileged, not privileged, or privileged with exceptions. The analysis requires considering multiple factors: the nature of the communication, the parties involved, the subject matter, the purpose of the communication, and any applicable exceptions like the crime-fraud

exception or common interest doctrine.

A reasoning model can work through this analysis methodically. Given an email chain with multiple participants, some in-house counsel and some business personnel, discussing both legal strategy and business operations, a reasoning model might:

- Identify each communication segment and determine who sent it to whom
- Analyze the subject matter of each segment separately
- Determine whether counsel was providing legal advice or business guidance
- Consider whether business participants on the email waive privilege
- Evaluate whether any exceptions apply
- Reach a reasoned conclusion with appropriate qualifications

This kind of multi-factor analysis is exactly what reasoning models do well. They can maintain logical consistency across the various considerations and avoid the kind of oversimplification that leads to errors in privilege review.

Analyzing Contradictory Testimony

Consider a case with testimony from fifteen witnesses about a series of meetings that occurred over six months. Some witnesses remember certain details clearly while others are vague. Testimony conflicts about who said what, when decisions were made, and what information various participants had at different times.

A reasoning model can systematically work through these contradictions. It might start by creating a timeline of undisputed facts, then layer in each witness's testimony about disputed events, identify where accounts diverge, consider the reliability of different witnesses based on their role and proximity to events, and synthesize a coherent narrative that accounts for the inconsistencies.

The model's ability to hold multiple competing accounts in working memory, evaluate them against independent evidence, and reason about credibility makes it particularly valuable for this kind of analysis. Rather than simply noting contradictions, it can work through possible explanations and evaluate which accounts are best supported by the full record.

Timeline Construction with Logical Inference

Building comprehensive chronologies from document sets often requires logical inference. A document might reference "the meeting last Tuesday" without specifying which Tuesday. Another might mention "the proposal we discussed" without naming the proposal. Constructing an accurate timeline requires piecing together these references through careful reasoning.

Reasoning models can work through these inference chains. They can identify references to the same event across different documents, use context clues to date undated materials, recognize when documents are responding to earlier communications, and flag timeline gaps that suggest missing documents. This kind of detective work, which requires holding multiple pieces of information in mind while working through logical connections, plays to reasoning models' strengths.

Document Authenticity Assessment

Sometimes you need to evaluate whether a document is what it purports to be. An email's metadata might not match its apparent content. Document properties might be inconsistent with the stated creation date. References in the document might conflict with other evidence about when it was created.

A reasoning model can systematically evaluate authenticity concerns. It might examine meta-data, compare document formatting and language with other documents from the same alleged time period, check whether referenced events had actually occurred by the document's stated date, identify anachronisms, and evaluate the overall pattern of evidence about when and how the document was created. This kind of forensic analysis requires exactly the sort of methodical, multi-factor reasoning that these models provide.

Evaluating Opposing Arguments

Reviewing opposing counsel's briefs, productions, or privilege logs often requires identifying logical gaps, unsupported assertions, or inconsistencies. A reasoning model can work through an opponent's argument step by step, test each premise against the evidence, identify logical leaps, spot selective use of facts, and craft responses that address the argument's actual weaknesses rather than straw men.

This application is particularly valuable when dealing with sophisticated opposition. The model's ability to engage with complex legal reasoning, maintain logical rigor, and identify subtle flaws in argumentation helps you respond more effectively than you might with quick-glance review.

Investigation Planning

At the start of an investigation, you often face choices about where to focus limited resources. Which custodians are likely to have relevant information? What time periods matter most? What search terms will be most productive? What patterns should you look for?

A reasoning model can help think through investigation strategy. Given the initial allegations and available information, it can reason about who would likely be involved, what kinds of evidence would be most probative, where you might find that evidence, and what sequence of investigative steps would be most efficient. This kind of strategic reasoning, which requires understanding the legal issues, anticipating what evidence might exist, and planning a logical investigation sequence, benefits from the model's ability to work through multiple considerations systematically.

Advanced Reasoning: Multi-Factor Relevance Analysis

Consider a trade secrets case where the plaintiff alleges the defendant hired away key employees and used them to develop competing products. Relevance depends on several factors: when the employees were hired, what knowledge they had, what they worked on after joining the defendant, whether they accessed old files, what the competitive products actually contain, and the timing of product development.

A standard LLM given a document about hiring practices might flag it as relevant because it mentions employee transfers. But a reasoning model can work through the full analysis. It might determine that while the document discusses employee transfers, it relates to a different department that had nothing to do with the technology at issue, involves transfers that occurred after the relevant time period, and therefore is not actually relevant despite superficial similarity to the case themes.

This kind of contextual reasoning, which requires understanding the legal theory, knowing what facts would actually matter to that theory, and distinguishing between superficially similar situations, represents exactly where reasoning models add value. They can avoid false positives that result from simplistic pattern matching and instead evaluate relevance based on actual legal and factual significance.

Cost and Speed Considerations

Reasoning models cost more and run slower than standard models. A task that takes a standard model five seconds might take a reasoning model thirty seconds or even minutes. Cost per query can be five to ten times higher. These tradeoffs mean you need to be strategic about when to use reasoning capabilities.

Use reasoning models for high-value analytical tasks where accuracy and thoroughness justify the extra cost and time. Privilege determinations on sensitive documents, analysis of key testimony, evaluation of critical evidence, and strategic planning benefit from the enhanced reasoning. Routine tasks like document categorization, simple summaries, or high-volume review are better suited to standard models.

Think of reasoning models as the discovery equivalent of bringing in senior counsel for the complex issues while letting junior attorneys handle routine matters. You would not have a partner review every document in a production, but you would want partner-level attention on the difficult privilege calls or the key contradictory evidence. Similarly, you would not run reasoning models on every task, but you would use them where careful analysis matters most.

Combining Reasoning Models with Other AI Capabilities

Reasoning models work well in combination with other AI capabilities. You might use standard models for initial document review to identify potentially relevant materials, then apply reasoning models to analyze the most important documents in depth. Or you might use agentic systems to gather evidence from multiple sources, then employ a reasoning model to synthesize that evidence into a coherent legal analysis.

One effective pattern is the two-stage approach: use fast, inexpensive models to filter and organize large volumes of information, then use reasoning models for the complex analytical tasks that require careful thought. This combines efficiency for routine tasks with thoroughness for difficult problems.

Another useful combination is having reasoning models verify the output of other AI systems. An agent might analyze thousands of documents and produce a report, then a reasoning model might review that report for logical consistency, check whether conclusions are properly supported, and identify any analytical gaps. This provides a form of automated quality control.

Limitations and Cautions

Reasoning models are not perfect. They can still hallucinate, though typically less than standard models. Their extended thinking process provides more opportunities to catch and correct errors, but it does not eliminate them. The same verification practices that apply to standard model output apply equally to reasoning models.

The thinking process, while more thorough than standard models, still differs from human reasoning in important ways. The models lack true understanding and cannot apply common sense in the way humans do. They can follow logical chains impressively well, but they can also pursue reasoning paths that seem logical but rest on faulty premises. Human oversight remains essential.

Reasoning models work best with clear, well-defined tasks. Vague instructions or ambiguous questions can lead them down unproductive analytical paths. When using reasoning models, invest time in crafting precise prompts that specify what analysis you want, what factors to consider, and what kind of conclusion you need.

The extended thinking time can sometimes be unproductive. The model might explore analytical paths that turn out to be irrelevant, or spend time on considerations that do not ultimately matter. This is part of the deliberation process, but it means that longer thinking time does not always equal better results. For simple questions, the extra time provides no benefit.

Best Practices for Using Reasoning Models

Several practices will help you get the most from reasoning models:

Be specific about what reasoning you want: Rather than asking a reasoning model to “analyze this document,” explain what kind of analysis matters. “Evaluate whether this email is protected by attorney-client privilege, considering the parties involved, the subject matter, and whether the primary purpose was to provide legal advice” gives the model clear analytical direction.

Provide relevant context: The model reasons better when it has the necessary background. Include relevant legal standards, case-specific facts, and any constraints on the analysis. If you are asking about privilege, provide your jurisdiction’s privilege standards rather than making the model guess.

Use for genuinely complex tasks: Simple yes/no questions or straightforward document summaries do not benefit from extended reasoning. Save reasoning models for tasks that actually require working through multiple considerations or complex logical chains.

Request explanations of reasoning: Ask the model to show its work. Understanding how it reached a conclusion helps you verify the analysis and identify any logical errors. Many reasoning models can provide summaries of their thinking process.

Verify conclusions: Even carefully reasoned analysis can be wrong. Check citations to source documents, verify that legal standards are applied correctly, and test conclusions against your own understanding of the facts and law.

Iterate on difficult problems: If a reasoning model's first analysis seems incomplete or misses important considerations, provide feedback and ask it to reconsider. The model can incorporate your guidance into a revised analysis.

The Future of Reasoning Models in Discovery

Reasoning capabilities are likely to improve rapidly. Current models already show impressive analytical abilities, but they will become more sophisticated at identifying relevant considerations, catching their own errors, and reasoning about complex legal and factual problems. We can expect faster reasoning with better cost-efficiency, making these capabilities practical for a broader range of discovery tasks.

Integration with specialized legal knowledge will make reasoning models more effective. Models trained on case law, discovery practices, and legal reasoning will better understand what matters in legal analysis and how to apply legal standards to facts. This specialized knowledge combined with general reasoning ability will create powerful tools for legal analysis.

We will see reasoning models integrated into complete discovery workflows. Rather than being separate tools you invoke for specific tasks, reasoning capabilities will become embedded in document review platforms, investigation systems, and case management tools. The software will recognize when a task requires careful analysis and automatically apply reasoning models where appropriate.

Perhaps most intriguingly, reasoning models will enable AI systems that can explain their decisions in ways that satisfy legal and professional standards. When a model makes a privilege determination or evaluates document relevance, it will be able to provide a clear, logical explanation of its reasoning that lawyers can review, critique, and defend if challenged. This transparency will make AI analysis more trustworthy and defensible.

Evolution

Reasoning models represent a significant advance in AI capabilities for discovery and investigation. Their ability to work through complex problems methodically, maintain logical consistency across multi-step analysis, and evaluate their own conclusions makes them valuable tools for the difficult analytical tasks that discovery professionals face daily.

These models are not replacements for human judgment. They are cognitive tools that extend our analytical capabilities, much as calculators extended our mathematical capabilities. Used thoughtfully, they help us analyze evidence more thoroughly, identify patterns we might miss, and work through complex logical problems more systematically. Used carelessly, they can lead us astray with confident-sounding but flawed analysis.

The key is knowing when and how to deploy reasoning capabilities. For routine tasks, standard models work fine. For complex analytical problems where logical rigor matters, reasoning models provide genuine value. As these capabilities mature and become more accessible, discovery professionals who learn to use them effectively will have significant advantages in handling complex investigations and litigation.



The introduction of reasoning models marks another step in AI's evolution from simple automation to genuine cognitive assistance. Discovery professionals who understand these capabilities and integrate them thoughtfully into their practice will be better equipped to handle the increasingly complex and data-intensive investigations that characterize modern litigation.

Looking Ahead

Reasoning models provide powerful analytical capabilities, but discovery work rarely involves isolated analytical tasks. Real investigations require coordinating multiple analysis steps: searching document collections, extracting relevant information, synthesizing findings across sources, generating work product, and iterating based on what you learn. A reasoning model can analyze a contract brilliantly, but someone still needs to find the right contracts, determine what questions to ask about each one, connect findings across multiple agreements, and produce a coherent analysis.

This coordination challenge is where agentic workflows become essential. In Chapter 8, we examine how AI systems can plan and execute complex discovery tasks with minimal supervision, combining reasoning capabilities with the ability to use tools, make decisions about next steps, and work toward defined goals. These systems don't just analyze better—they work differently, handling entire investigation workflows that previously required extensive human orchestration.

08

Agentic Workflows in Discovery

—

CHAPTER EIGHT

Agentic Workflows in Discovery

In Chapter 6, we explored the components and architectures that make AI systems “agentic”—capable of planning, executing multi-step workflows, and iterating toward goals autonomously. Now we examine how these abstract capabilities translate into practical discovery applications that transform how legal professionals conduct investigations and analyze evidence.

The workflows we explore in this chapter go beyond simple prompt-response interactions. They represent sophisticated investigative processes where AI systems autonomously pursue complex objectives, make strategic decisions, adapt to emerging information, and produce comprehensive work product that previously required days or weeks of human effort.

The workflows we explore represent a spectrum of capabilities. Some, like autonomous document analysis and timeline construction, are beginning to appear in production systems. Others, particularly sophisticated multi-agent coordination and real-time deposition support, represent near-term possibilities that current technology could support but that haven’t yet been fully realized in commercial products. Understanding this full range prepares you to evaluate emerging vendor offerings and envision how discovery work will evolve.

The workflows in this chapter illustrate the range of what agentic AI makes possible. Some capabilities—like autonomous document analysis following multiple search iterations—are entering the market now. Others, particularly sophisticated multi-agent systems and real-time deposition support, represent where the technology is heading over the next 2-3 years. Understanding the full spectrum helps you recognize emerging capabilities as vendors roll them out and envision how your practice will evolve.

Document Analysis Agents

Document analysis represents one of the most promising applications of agentic AI in discovery, with early implementations beginning to appear in commercial systems. Traditional document review follows linear patterns: search for documents, review results, refine searches, repeat. Agentic document analysis agents transform this into autonomous investigations that search strategically, evaluate results critically, and iterate until they achieve comprehensive understanding.

The Traditional Document Analysis Challenge

Discovery professionals routinely face scenarios like this:

A partner asks: “We need to understand what the board knew about the product defect and when they knew it. Review the document production and give me a comprehensive report.”

Traditional workflow requires:

- Developing keyword searches for board-related documents
- Reviewing search results to identify relevant materials
- Reading board minutes, emails, and reports
- Building timeline of board knowledge
- Identifying gaps requiring additional searches
- Conducting supplemental searches
- Synthesizing findings across sources
- Writing comprehensive report

This process might consume several days of attorney or senior paralegal time, involving iterative searching, extensive reading, and careful synthesis. Each iteration requires human decision-making about next steps.

How Document Analysis Agents Work

An agentic document analysis system would approach this same task autonomously. While the level of sophistication varies across current implementations, the workflow demonstrates the technology's potential:

Phase 1: Goal Understanding and Initial Planning

The agent receives the goal: "Understand what the board knew about the product defect and when they knew it."

The agent analyzes this goal, identifying key requirements:

- Subject matter: product defect information
- Key actors: board members
- Temporal component: when knowledge was acquired
- Deliverable: comprehensive report with timeline

The agent develops an initial strategic plan:

1. Identify who served on the board during relevant period
2. Search for board-related documents (minutes, presentations, reports)
3. Search for documents about the product defect
4. Identify intersection of board and defect documents
5. Analyze these documents for evidence of board knowledge
6. Build timeline of when information reached board
7. Identify key documents requiring detailed analysis
8. Search for related communications explaining or following up on board discussions
9. Assess completeness and identify gaps
10. Conduct targeted searches to address gaps
11. Synthesize findings into comprehensive narrative
12. Generate timeline visualization
13. Compile report with citations to source documents

Phase 2: Information Gathering and Analysis

The agent begins executing its plan, but unlike a human following a checklist, it adapts continuously based on what it discovers.

Initial board identification: The agent searches for documents mentioning “board of directors,” “board meeting,” “board members.” It analyzes results to extract names of individuals serving on the board, compiling a list of key actors.

Strategic search execution: Rather than executing a single search, the agent orchestrates multiple search strategies simultaneously or sequentially:

- Semantic search for documents conceptually about product defects
- Boolean searches for specific technical terms about the defect
- Custodian-based searches focusing on board members identified
- Date-restricted searches around periods when defect was reported
- Email thread searches to capture board communications

The agent doesn’t just execute these searches—it evaluates results. After each search, it assesses:

- Are results relevant to the goal?
- What new information do results reveal?
- What patterns emerge?
- What’s still missing?
- What should be searched next?

Adaptive investigation: As the agent analyzes initial results, it discovers information that modifies its approach. Perhaps it finds:

- A critical email from an engineer to a board member six months earlier than expected
- References to a committee studying the issue (requiring committee searches)
- Technical reports presented to the board (requiring searches for those specific reports)
- Code names or project names used internally (requiring searches with those terms)

The agent dynamically adjusts its plan to pursue these leads:

Updated Plan:

- Original items 1-5: Completed
- NEW: Search for emails from Engineer X to board members 6+ months earlier
- NEW: Identify and search Safety Committee documents
- NEW: Search for “Project Phoenix” (code name discovered)
- NEW: Search for technical reports cited in board presentations
- Continue with original items 6-13, modified based on new findings

Phase 3: Deep Analysis and Synthesis

As relevant documents accumulate, the agent performs sophisticated analysis:

Document clustering: The agent groups related documents—email threads, meeting minutes with corresponding presentations, follow-up communications, etc. This clustering reveals relationships humans might miss across large document populations.

Timeline construction: The agent extracts temporal information from documents, building a chronology of when information about the defect reached various board members, when board meetings discussed it, and when decisions were made. This timeline identifies critical periods and potential gaps.

Knowledge attribution: For each piece of defect information, the agent tracks:

- When did it first appear in documents?
- Which board members received it?
- In what context (board meeting, email, report)?
- What actions followed?
- Were there later contradictory statements?

Gap identification: The agent identifies information gaps that additional searching might fill:

- Board meetings without accessible minutes
- Time periods with no relevant communications
- Board members not appearing in documents when they should
- Technical details mentioned but not explained in available documents
- References to documents not found in production

Phase 4: Targeted Gap-Filling

Based on gap identification, the agent conducts targeted follow-up investigations:

Precision searches: Rather than broad searches, the agent executes highly specific searches designed to find particular missing information:

- “Board meeting” AND “March 15” (to find minutes from specific meeting)
- Custodian:”Board Member X” AND date range (to find that member’s communications during silent period)
- Attachment searches for technical reports referenced but not found

Relationship exploration: The agent explores relationships between documents and individuals:

- Who corresponded with whom about the defect?
- Which external parties provided information to board members?
- What documents were attachments to communications analyzed?

Iterative refinement: If targeted searches don’t fill gaps, the agent tries alternative approaches:

- Broader date ranges
- Related terms or synonyms
- Communications mentioning the missing documents

- Correspondence with administrative staff who might have relevant information

Phase 5: Synthesis and Reporting

After thorough investigation, the agent synthesizes findings:

Narrative construction: The agent builds a coherent narrative answering the original question:

- What defect information existed and when it was created?
- How that information reached board members?
- What board discussions occurred?
- What decisions or actions board took?
- Evolution of board knowledge over time?
- Any contradictions or inconsistencies in the record?

Evidence support: Every substantive claim in the narrative includes citations to specific documents with IDs allowing immediate verification:

“On January 15, the board received a preliminary report indicating potential defects in the product’s safety mechanism (Doc ID 894523). The report noted that testing had revealed failure rates of 2.3% under stress conditions (Doc ID 894523, p. 7).”

Timeline generation: The agent creates a detailed timeline showing key events, knowledge milestones, and decision points, each linked to supporting documents.

Completeness assessment: The agent evaluates its own work:

- Were all aspects of the goal addressed?
- Are there remaining gaps or uncertainties?
- What information proved impossible to find?
- What alternative interpretations exist?

Final report compilation: The agent produces a comprehensive report including:

- Executive summary answering the original question
- Detailed narrative with full citations
- Timeline visualization
- List of key documents
- Methodology description
- Identified gaps or limitations
- Recommendations for further investigation if needed

Advantages Over Traditional Approaches

As these systems mature, agentic workflows will provide several advantages:

- 1. Comprehensiveness:** The agent systematically pursues the question from multiple angles, less likely to miss relevant documents than human reviewers who might not think

of all search strategies.

2. **Consistency:** The agent applies consistent analytical criteria across all documents, avoiding fatigue or attention lapses that affect human review.
3. **Speed:** What might require days of human work completes in hours or less, with the agent working continuously without breaks.
4. **Documentation:** The agent's process creates a complete audit trail of search strategies, decision points, and analytical steps, supporting defensibility.
5. **Adaptability:** The agent dynamically adjusts its approach based on discoveries, pursuing leads that emerge during investigation.
6. **Scale:** The agent can process document volumes that would overwhelm human reviewers, maintaining quality across large populations.

While an agentic system might work for hours building this comprehensive analysis, your professional responsibility doesn't disappear. The output requires verification—sampling the cited documents to confirm the agent correctly interpreted them, checking that the timeline logic makes sense, and applying your judgment about whether gaps identified are meaningful or routine. The agent dramatically reduces the time to gather and organize evidence, but strategic assessment remains your job

Investigation Agents: Building Timelines and Mapping Relationships

Beyond analyzing documents to answer specific questions, agentic systems are being developed to conduct broader investigations that construct timelines, map relationships between entities, and identify patterns across large document populations. Some of these capabilities exist in production systems today, while others represent near-term development targets.

Timeline Construction Agents

Chronology proves critical in many investigations. What happened when? Who knew what at each point? How did events unfold? Constructing comprehensive timelines from scattered documents traditionally requires extensive manual work. Timeline construction agents automate this process while maintaining accuracy.

How timeline agents work:

The agent searches across the document population to identify relevant materials, then extracts temporal information from each document—not just document dates, but dates mentioned in content, temporal relationships between events, and even estimated timeframes when precise dates aren't stated.

From this temporal data, the agent identifies discrete events worthy of timeline inclusion: meetings, decisions, regulatory filings, personnel changes, and other significant occurrences. For each event, it associates relevant entities (people, organizations, products) and supporting documents. The agent then maps relationships between events, identifying causal connec-

tions, sequential dependencies, and contradictions that require reconciliation.

Critically, the agent identifies temporal gaps that suggest missing information—long periods without documented activity, events that should have intermediate steps, or referenced communications without documentation. These gaps guide targeted follow-up searches.

The agent produces timeline outputs in multiple formats: chronological narratives with document citations, visual timelines, sortable event tables, and network diagrams showing relationships between events.

Example application:

Consider an investigation into events leading to a product recall. The agent searches broadly, identifies thousands of relevant documents, and extracts dates and events from each. It identifies key events spanning product design decisions through customer complaints, internal investigations, executive discussions, and regulatory communications.

The agent maps how customer complaints triggered an internal investigation, which revealed a manufacturing issue traceable to a specific process change decided by particular executives. It identifies critical gaps—a two-month period between first complaints and internal investigation, missing documentation of executive discussions before the recall decision, and absent quality control reports during the critical period. Based on these gaps, it conducts targeted searches to find missing information.

The result: a comprehensive timeline with 200+ events, each linked to supporting documents, along with visual representations highlighting key decision points and flags for events requiring additional investigation.

Value for discovery:

Timeline agents—in current and emerging forms—process document volumes humans cannot manually review, identify temporal patterns across scattered materials, and find gaps suggesting missing or withheld documents. They support both case strategy and depositions while creating demonstrative aids for hearings or trial. The automation doesn't replace professional judgment—it provides comprehensive chronological frameworks that legal professionals can analyze, challenge, and refine.

Relationship Mapping Agents

Understanding who communicated with whom about what, which organizations had business relationships, and how entities connect reveals patterns critical to investigations. Relationship mapping agents extract and visualize these connections.

How relationship mapping agents work:

The agent extracts entities from document populations—people, organizations, products, locations, and concepts—then identifies relationships between them. These relationships span multiple dimensions: communications patterns, organizational affiliations, business partnerships, knowledge sharing, spatial connections, and conceptual links around shared topics.

For each relationship, the agent characterizes its nature: strength based on interaction frequency, directionality (one-way versus reciprocal), temporal patterns showing when relationships existed, context distinguishing business from personal connections, topics discussed, and even sentiment or tone. The agent builds network representations where nodes represent entities and edges represent relationships, with edge weights reflecting relationship strength and temporal layers showing evolution over time.

The agent identifies significant patterns—central actors who are highly connected, clusters of entities with dense internal connections, and bridges connecting otherwise separate groups. It maps information flow paths and flags unusual relationships or unexpected connections. Critically, it performs anomaly detection, identifying communications outside expected patterns, connections between entities that shouldn't interact, sudden relationship intensity changes, individuals apparently avoiding documentation, and information silos.

Example application:

Consider mapping relationships among executives, board members, and consultants during merger negotiations. The agent extracts 147 individuals and 23 organizations from the document production, then analyzes communications to identify relationship patterns. It discovers that the CFO had frequent communications with a competing bidder, an external consultant appeared in many board discussions despite little documentation of retention, separate communication streams existed between board committees and management, key strategic decisions were made via personal rather than corporate email accounts, and one board member never appeared in any merger-related documents.

The agent maps the communication network, revealing primary participants and interaction frequencies, information flow from legal counsel through the CEO to the board, side channels bypassing formal processes, and external parties receiving internal information. It flags the CFO's competitor communications, the undocumented consultant role, the absent board member, and the use of personal accounts for corporate decisions—each warranting investigation. The deliverable includes visual network diagrams and targeted search recommendations for follow-up.

Value for discovery:

Relationship mapping agents reveal hidden connections among parties, identify individuals with knowledge relevant to depositions, expose information flows and potential leaks, and support conspiracy or fraud investigations. They process communication volumes that would overwhelm manual analysis, identifying patterns invisible in small samples. The resulting network visualizations guide strategic investigation priorities and create compelling demonstrative evidence for hearings or trial.

Real-Time Deposition Support

One of the most ambitious agentic applications under development is real-time support during depositions, where attorneys need immediate access to testimony, documents, and case information. While elements of this capability exist today—such as transcript search and document retrieval—the integrated system described here represents a near-term development target rather than a current product.

The Deposition Challenge

During depositions, attorneys must listen carefully to testimony, formulate follow-up questions, identify contradictions with prior statements, recall relevant documents, adjust examination strategy dynamically, and take detailed notes. This cognitive load makes it difficult to simultaneously search case materials for supporting documents or prior testimony. Associates often sit beside examining attorneys to search documents, but even skilled associates can't search fast enough to keep pace with dynamic examinations.

Agentic Deposition Assistant

An agentic deposition assistant would provide immediate support throughout the deposition process. Before the deposition, the agent analyzes all prior testimony from the witness, reviews associated documents, identifies key topics likely to arise, and builds an anticipatory search index. During the deposition, it processes the transcript in real-time via court reporter live feed or audio-to-text, identifies key claims or statements, automatically searches for related documents, flags contradictions with prior testimony, and surfaces relevant documents proactively.

The dynamic assistance works like this: When a witness claims "I never saw that report before March," the agent immediately searches for the report, finds an email from January showing the witness received it, alerts the attorney to the contradiction, and displays the email with relevant portions highlighted. When a witness mentions "We discussed that in the board meeting," the agent retrieves board meeting minutes, identifies the relevant discussion, summarizes what the minutes say, and flags discrepancies between testimony and minutes. When an attorney asks about "communications with Acme Corp," the agent pulls all Acme Corp communications, summarizes key themes, identifies documents not yet covered in questioning, and suggests follow-up questions.

After the deposition, the agent generates a detailed summary of testimony, identifies all contradictions with documents or prior testimony, creates an issue-spotting memo highlighting significant testimony, updates the case knowledge base with new information learned, and generates recommended document requests based on testimony gaps.

Implementation Requirements

Effective deposition agents will require low-latency operation with responses arriving within seconds, high accuracy to avoid disrupting examination flow with false contradictions or irrelevant documents, unobtrusive interfaces delivering information via tablet or laptop without distracting the attorney from the witness, attorney control allowing explicit queries or autonomous operation, and confidence signaling so attorneys can judge the reliability of findings.

Multi-Agent Investigation Systems

For complex investigations requiring diverse analytical approaches, multi-agent systems represent an emerging architecture where specialized agents collaborate under coordination frameworks. While research prototypes demonstrate this approach, commercial implementations are still developing.

Rather than one generalist agent handling everything, a sophisticated multi-agent discovery system deploys specialized agents for search strategy, document classification, content analysis, timeline building, relationship mapping, synthesis, and quality assurance. Each agent excels at its specific role, leading to higher quality results than a single agent could achieve. Multiple agents work simultaneously, dramatically reducing completion time while maintaining quality through verification processes.

The multi-agent example that follows illustrates how current technology could be orchestrated to handle complex discovery tasks. Consider this a roadmap for where the technology is heading rather than a description of widely available products.

Example: Multi-Agent Product Liability Investigation

This example demonstrates how specialized agents could collaborate in a mature implementation to comprehensively analyze documents related to product defect allegations.

The investigation begins with a Search Strategist Agent developing search strategies covering product names, defect descriptions, customer complaints, engineering documents, quality control, and executive communications—yielding 50,000 potentially relevant documents. A Document Classifier Agent categorizes these into customer communications, internal emails, engineering documents, quality control reports, executive materials, and marginally relevant documents, routing high-priority categories forward while flagging potentially privileged documents for human review.

In the detailed analysis phase, a Content Analyzer Agent reads the high-priority documents, extracting key information about defect awareness, engineering responses, and executive decisions while generating summaries and identifying documents requiring deep analysis. Simultaneously, a Timeline Builder Agent processes documents to construct a chronology showing first customer complaints 18 months before recall, engineering investigation initiation, defect confirmation in testing, executive presentations, regulatory notification, and recall announcement—identifying significant timeline gaps requiring investigation. A Relationship Mapper Agent analyzes communications to map information flow from customer service through engineering to executives, identifying which executives received what information when, flagging concerning patterns like executives avoiding documentation.

Based on gaps and anomalies identified by the Timeline and Relationship agents, the Search Strategist Agent conducts a second pass with targeted searches for missing engineering reports, executive communications during suspicious periods, and external consultant correspondence. The Content Analyzer performs detailed analysis of newly discovered documents, cross-referencing with existing understanding and identifying contradictions requiring resolution.

A Synthesis Agent combines findings from all specialized agents into a comprehensive narrative addressing what defect existed and when it was discovered, how information flowed through the organization, what decisions were made and by whom, whether responses were adequate and timely, and evidence of knowledge or intent. It identifies key documents and witnesses for litigation and develops theories of liability and defense. A Quality Assurance Agent reviews this narrative against source documents, verifies that citations support claims, checks timeline accuracy, identifies unsupported conclusions, and recommends additional investigation where needed. A Coordination Agent compiles final deliverables: executive summary, detailed

investigative report with citations, timeline with all events, relationship network diagrams, key document list, witness list with knowledge summaries, and recommendations for depositions and discovery requests.

Finally, a senior attorney reviews the multi-agent investigation results, verifies key findings against source documents, identifies areas needing additional investigation, and provides feedback. Based on this feedback, the Coordination Agent directs specialized agents to conduct follow-up investigation, repeating the process until the attorney determines the investigation is sufficiently complete.

Advantages of Multi-Agent Systems

When fully realized, multi-agent architectures would provide significant benefits. Specialization allows each agent optimized for its specific task to perform better than generalist approaches. Parallelization enables multiple agents to work simultaneously, dramatically reducing completion time. Quality assurance through dedicated QA agents catches errors before human review. The architecture offers scalability (additional agents can be added without redesigning the system), modularity (individual agents can be updated or replaced independently), and robustness (if one agent fails, others continue working and the system degrades gracefully rather than failing completely).

Understanding the Economics

Multi-agent investigations running for hours across hundreds of thousands of documents consume significant computational resources. While the cost per document may be lower than manual review, expect these comprehensive investigations to cost thousands of dollars in processing fees. The value proposition lies in getting depth and breadth of analysis impossible through traditional methods at any reasonable price—not in getting traditional analysis cheaper.

What Requires Your Oversight?

Agentic systems can pursue unproductive directions, misinterpret document relationships, or miss obvious patterns that context would reveal. Timeline construction might miss informal communications or misdate events mentioned vaguely. Relationship mapping might identify spurious connections while missing significant ones. Quality assurance agents catch some errors but not all.

Your oversight focuses on three questions: Does this make sense given what I know about this case? Are there obvious gaps or misinterpretations? What assumptions did the agent make that I need to verify? The agent's comprehensive analysis provides the foundation, but your judgment and experience remain essential."

Looking Ahead

This chapter demonstrated how agentic AI capabilities translate into practical discovery workflows. The examples span a range of development stages: document analysis agents with basic autonomous capabilities are entering the market now, timeline construction and relationship



mapping tools exist in limited forms, while sophisticated multi-agent investigation systems remain primarily in development. Together, they illustrate how discovery work will transform as these technologies mature.

As these capabilities develop, they won't replace human professionals—they'll augment human expertise, handling volume and routine analysis while legal professionals focus on strategy, judgment, and critical decision-making. The emerging result is discovery work that's faster, more comprehensive, and more consistent while remaining under professional oversight.

Chapter 9 addresses the critical question of how to implement and control these systems effectively. Agentic AI's power brings risks—autonomous systems can pursue unproductive directions, consume excessive resources, or produce outputs requiring extensive verification. The next chapter provides frameworks for deploying agentic systems with appropriate controls, ensuring they serve professional needs without creating unacceptable risks.

We'll examine when agentic approaches are appropriate versus when simpler methods suffice, how to set boundaries and guardrails, how to monitor agent behavior, and how to maintain human oversight while preserving efficiency benefits. These implementation considerations determine whether agentic AI becomes a valuable tool or an expensive liability in your practice.

09

Ethical Considerations for Agentic Systems

CHAPTER NINE

Ethical Considerations for Agentic Systems

Chapter 5 established the general ethical framework for using Generative AI in legal practice—the duty of competence, confidentiality obligations, verification requirements, transparency expectations, and billing practices. Those principles apply to all AI use in legal work. But agentic systems raise additional ethical questions that warrant specific attention.

Agentic systems execute multi-step investigations between checkpoints. This improves efficiency but requires clear boundaries, auditable operations, and human control over strategy and conclusions. This chapter examines the ethical considerations unique to agentic AI—how increased autonomy affects professional responsibilities, what transparency obligations arise when systems operate autonomously, who bears responsibility when agents make mistakes, how to ensure quality and reliability standards are met, what limits should be placed on agent autonomy, and how to maintain the professional judgment that defines legal practice.

Autonomy and Professional Responsibility

The defining characteristic of agentic systems is their autonomy—the ability to pursue goals through self-directed actions rather than executing prescribed steps. This autonomy creates tension with professional responsibility rules premised on lawyers maintaining direct control over legal work.

The Control Question

Traditional professional responsibility rules assume lawyers directly control or closely supervise all work performed on behalf of clients. When a lawyer delegates work to an associate or paralegal, the lawyer maintains oversight through assignment instructions, interim reviews, and final approval. The supervising lawyer can intervene at any point, redirect the work, or reject outputs that don't meet standards.

Agentic systems operate differently. Once launched toward a goal, they make numerous operational decisions autonomously—which documents to retrieve, what search strategies to pursue, how to analyze findings, when to iterate versus moving forward. While humans establish goals and boundaries at the outset and review outputs at checkpoints, the agent operates independently between these touchpoints.

This raises the fundamental question: Does autonomous agent operation constitute improper delegation of professional responsibility? The answer depends on understanding what agents actually control versus what remains under human direction.

What Lawyers Control

Even with agentic systems, lawyers maintain control over all substantive and strategic deci-

sions. Lawyers decide what investigations to pursue, what goals to establish, what boundaries to set, what resources to authorize, when to intervene based on checkpoint reviews, what conclusions are supported by evidence, how to present findings to clients or courts, and what actions to recommend. Agents operate within the framework established by these human decisions.

The autonomy agents exercise is operational, not substantive. Agents decide how to search efficiently within defined scope and which documents to prioritize but don't determine investigation strategy or what conclusions follow from findings.

This distinction matters ethically. Professional responsibility rules require lawyers to exercise independent professional judgment on substantive matters affecting client interests. Those rules don't require lawyers to personally perform every operational task. The question is whether the lawyer maintains appropriate oversight and makes all substantive decisions.

Appropriate Delegation Boundaries

Agentic systems remain within appropriate delegation boundaries when lawyers establish clear investigation goals reflecting professional judgment about what information clients need, set boundaries ensuring agents operate within authorized scope, implement checkpoints providing opportunities to review progress and redirect if needed, verify agent outputs meet quality standards before use, and make all substantive decisions about conclusions and recommendations.

Agentic systems cross into inappropriate delegation when agents make strategic decisions about investigation scope or direction, determine what conclusions are supported by evidence without human verification, communicate findings or recommendations directly to clients without lawyer review, or make decisions involving professional judgment or ethical considerations.

The ethical line is that agents can autonomously execute work within parameters established by lawyers, but all decisions requiring professional judgment remain with human lawyers who bear ultimate responsibility.

Transparency and Disclosure Obligations

Transparency enables informed client consent, supports verification and accountability, facilitates oversight, and maintains trust in the legal system. Agentic systems raise particular challenges because their autonomous operation makes it harder to explain exactly what was done and why.

Transparency to Clients

Lawyers have ethical obligations to keep clients informed about significant developments in their matters and to explain matters to the extent necessary to permit clients to make informed decisions. These obligations apply when using agentic AI to perform work for clients.

The threshold question is whether AI use constitutes a "significant development" requiring client notification. The answer depends on the context. When agentic AI is used for routine work

where the method employed doesn't affect strategy or outcomes in ways clients should know about, notification may be unnecessary. When AI use involves novel methods, affects cost or timing significantly, or raises questions clients might reasonably want input on, disclosure becomes appropriate.

Best practice points toward proactive transparency. Many firms now include language in engagement letters indicating that AI tools may be used in providing legal services, subject to appropriate safeguards and oversight. This general disclosure establishes client awareness and consent without requiring notification for each specific instance.

When agentic systems produce findings or recommendations that will significantly affect client decisions, lawyers should explain the methodology. This doesn't require technical details about LLM architecture—it requires explaining in accessible terms what was done, how reliability was ensured, and what the findings mean.

Transparency to Courts and Opposing Counsel

Courts have inherent authority to inquire about methods used in litigation, and procedural rules may require disclosure in specific contexts like discovery disputes or expert testimony. When methodology becomes material—perhaps because opposing counsel questions completeness of review or because a discovery dispute raises questions about methods used—transparency becomes essential. Lawyers should not misrepresent AI-assisted work as entirely human-performed, nor should they obscure material information about methods when disclosure is relevant to assessing reliability or completeness.

Transparency About Limitations and Uncertainties

Most importantly, lawyers must be transparent about limitations and uncertainties in AI-assisted work. When agentic systems report that findings are based on limited information, when verification reveals potential gaps, or when results contain uncertainties, these limitations must be communicated appropriately. This transparency serves both ethical obligations and strategic interests. Clients deserve to know when conclusions rest on uncertain foundations.

Courts deserve to understand limitations in evidence or arguments presented to them. Practically, this means preserving and communicating the caveats and qualifications that should appear in agent outputs. When agents report “based on available documents” or “three sources provide conflicting information,” those qualifications should carry through to how findings are presented to clients and others.

Accountability and Error Attribution

When agentic systems make mistakes, questions arise about who bears responsibility and how accountability operates. These questions have both practical and ethical dimensions.

The Lawyer's Ultimate Responsibility

The foundational principle is clear: lawyers remain professionally responsible for all work product provided to clients, regardless of what tools or assistants contributed to that work.

This applies to work performed by associates, paralegals, vendors, AI systems, or any other contributor. The lawyer's professional responsibility is non-delegable.

When an agentic system produces inaccurate findings, misrepresents source documents, overlooks critical information, or makes other errors, the lawyer using that system bears professional responsibility. The lawyer cannot deflect responsibility by claiming the AI made the mistake. This principle is well established in other delegation contexts. When an associate makes a research error, the supervising partner remains responsible. The same applies to AI tools—they're tools the lawyer chose to use, and the lawyer remains responsible for outputs generated with those tools.

Comparative Fault and Mitigation

While ultimate responsibility rests with lawyers, practical questions of comparative fault and mitigation remain relevant. When a lawyer took all reasonable precautions but an AI system failed in unpredictable ways, disciplinary authorities may view the situation differently than when a lawyer deployed AI recklessly without safeguards.

Taking reasonable precautions doesn't eliminate lawyer responsibility but may mitigate consequences. A lawyer who implemented robust verification protocols, conducted appropriate sampling, and had reasonable systems for catching errors has exercised better judgment than a lawyer who blindly relied on AI outputs without verification—even if errors occurred in both cases. This creates practical incentives for robust quality assurance systems.

Learning from Errors

Accountability includes learning from mistakes to prevent recurrence. When agentic systems produce errors, organizations should investigate root causes, assess whether verification protocols should be enhanced, and determine whether guardrails should be adjusted.

Documentation plays an important role here. Maintaining records of errors identified, investigations conducted, improvements implemented, and lessons learned creates evidence of responsible oversight and continuous improvement.

Quality and Reliability Standards

Agentic systems must meet quality and reliability standards appropriate for legal work. These standards go beyond avoiding outright errors to ensuring outputs reliably meet professional standards.

Defining Adequate Quality

The quality threshold for agentic outputs should match what would be expected from human-performed work in similar contexts. For high-stakes litigation work where accuracy is critical, agentic outputs should meet the same stringent standards expected from senior attorneys. For preliminary internal investigations where some error tolerance exists, standards might be more relaxed—but not so relaxed that outputs are unreliable or misleading.

This situational approach recognizes that legal work involves varying stakes and quality requirements. Not every task requires perfect accuracy—some preliminary work accepts higher error rates as the trade-off for speed and efficiency. But even preliminary work must meet minimum reliability thresholds. Outputs that are more likely wrong than right, or that contain so many errors that they mislead rather than inform, fail to meet professional standards regardless of their preliminary nature.

Practically, this means calibrating agentic systems to task requirements. High-stakes work requires more stringent guardrails, more intensive verification, more conservative confidence thresholds, and more frequent human checkpoints. Lower-stakes work can accept somewhat lighter oversight—but never to the point where quality falls below acceptable minimums.

Reliability Over Time

Quality standards include reliability—the expectation that systems produce consistent results over time rather than alternating between excellent and poor performance unpredictably. Unreliable systems create risks even if average performance is acceptable, because users can’t distinguish when to trust outputs versus when to be skeptical.

This requires ongoing monitoring of system performance. Initial validation showing a system performs well doesn’t ensure continued reliability. Systems may degrade over time as underlying models are updated, as document collections change character, or as edge cases emerge that weren’t present during initial testing. Regular quality checks ensure continued reliability and identify when performance degrades requiring intervention.

When reliability problems emerge, organizations must respond appropriately. Continued use of systems known to produce unreliable outputs violates professional obligations. The response might involve enhanced verification protocols, adjusted guardrails, more intensive human oversight, or discontinuing use for particular applications until reliability issues are resolved.

Comparative Reliability

An important consideration is how agentic system reliability compares to alternatives. Agentic systems need not be perfect to be ethically usable—they need to be as reliable as or more reliable than the alternatives they replace, considering both accuracy and efficiency. If human review of thousands of documents produces certain error rates and patterns, agentic review should aim for comparable or better accuracy while offering efficiency advantages. The standard is whether agentic approaches meet professional quality expectations considering the full context of alternatives, stakes, and trade-offs.

Autonomy Limits and Human Judgment

Determining appropriate limits on agent autonomy requires balancing efficiency benefits against risks of excessive delegation. Some decisions inherently require human judgment and should never be delegated to autonomous systems.

Decisions Requiring Human Judgment

Certain categories of decisions must remain exclusively in human hands because they involve professional judgment, ethical considerations, or client interests that machines cannot appropriately weigh.

Strategic decisions about case direction, investigation scope, resource allocation, and litigation approaches require human judgment about client interests, risk assessment, and professional experience. Agentic systems can inform these decisions by gathering and analyzing information, but the decisions themselves must be made by lawyers exercising professional judgment.

Ethical decisions involving privilege, conflicts, disclosure obligations, billing practices, or professional conduct require applying ethical rules and professional judgment to specific facts. These decisions implicate lawyer responsibilities that cannot be delegated to autonomous systems. When agents encounter potential privilege issues, possible conflicts, or uncertain ethical questions, they must escalate to humans rather than attempting to resolve these matters autonomously.

Client counseling about strategy, settlement, risk assessment, or important decisions requires the attorney-client relationship and human understanding of client interests, values, and objectives. Agents should not communicate directly with clients about strategic matters or provide advice without human lawyer review and approval.

Novel situations without clear precedent require human judgment to navigate uncertainty and ambiguity. When agents encounter unprecedented patterns, unusual facts, or situations that don't fit established frameworks, they should escalate to humans who can apply judgment, creativity, and experience to novel problems.

Operational Autonomy Within Boundaries

Within appropriate boundaries, operational autonomy creates value without raising ethical concerns. Agents can autonomously decide which search strategies to use within validated methods, which documents to prioritize among relevant candidates, what level of detail to apply to different documents, how to structure outputs for clarity, when to iterate on searches versus moving forward, and what tools to use for particular tasks.

These operational decisions implement the strategy lawyers established but don't determine what that strategy should be. They involve technical optimization rather than professional judgment. Allowing agents to make these decisions autonomously captures efficiency benefits while preserving human control over matters requiring professional judgment.

The key is ensuring boundaries are clear, preventing agents from drifting into decisions requiring human judgment, implementing escalation mechanisms for edge cases, and maintaining human oversight at strategic points.

The Right Level of Autonomy

Finding the right autonomy level requires experimentation and adjustment. Too little autonomy eliminates the efficiency benefits justifying agentic approaches—if humans must approve

every operational decision, the system becomes too cumbersome to provide value. Too much autonomy creates risks of inappropriate delegation and loss of necessary control.

Organizations should start with relatively limited autonomy, gradually expanding as experience builds confidence and as verification demonstrates that agents respect boundaries and make appropriate operational decisions. This graduated approach allows learning what autonomy levels work well for different tasks and contexts while managing risk of excessive delegation.

Regular review of what decisions agents are making autonomously, whether those decisions reflect appropriate operational choices, whether agents are properly escalating decisions requiring human judgment, and whether the balance between autonomy and control is appropriate for each application guides adjustments over time.

Competence and the Duty to Understand

The fundamental ethical obligations for AI use—competence, independent judgment, confidentiality, and client consent—were examined in Chapter 5. Those principles apply fully to agentic systems, but the autonomous, multi-step nature of these systems creates particular challenges in meeting those obligations.

Professional responsibility rules require lawyers to maintain competence in technologies they use. For agentic systems, this creates particular challenges because the systems are complex and their operation can be opaque.

What Lawyers Should Understand

Lawyers using agentic systems need to understand how autonomous workflows can propagate errors across multiple steps, when and how to intervene in multi-step processes without undermining their value, what verification is necessary when you cannot easily trace how conclusions were reached, and what constraints and guard rails are appropriate for different applications. This understanding should be practical—knowing enough to deploy systems appropriately, recognize when problems may exist, and explain in accessible terms what was done and why it's reliable.

Building Competence

Building competence with agentic systems requires supervised use where experienced reviewers assess whether users are setting appropriate parameters, recognizing when to intervene in autonomous workflows, and implementing adequate verification of multi-step outputs. Over time, users develop intuitions about when agents are working well versus when autonomous processes need human direction. Competence includes recognizing when expertise beyond one's own is needed—when to consult technical experts about system behavior, when specialized verification is required for complex workflows, and when applications require additional oversight.

Preserving Professional Independence

Professional responsibility rules require lawyers to exercise independent professional judg-

ment on behalf of clients. With agentic systems, a significant risk is over-reliance—deferring to polished, confident-sounding outputs from multi-step autonomous processes rather than exercising independent judgment. When an agent has executed a complex workflow and produced comprehensive analysis, the temptation is to accept it without adequate critical evaluation.

Professional judgment must remain the final arbiter of conclusions reached, strategies pursued, and work product delivered to clients. Agentic systems can gather and analyze information through autonomous workflows, but the judgments themselves must be made by lawyers based on their evaluation of all relevant factors. This means sometimes disagreeing with agent outputs even when they result from sophisticated multi-step analysis. If conclusions don't seem supported by evidence, if important information appears to have been missed, or if suggested approaches seem inappropriate, professional judgment must prevail. Professional responsibility doesn't require that lawyer judgment always be correct—it requires that important judgments be made by lawyers exercising genuine professional judgment rather than abdicating decisions to automated systems.

Client Consent and Informed Decision-Making

Clients have rights to make informed decisions about their legal representation, including what methods are used. The consent principles from Chapter 5 apply to agentic systems, but the autonomous nature of these systems may raise additional client interest or concern.

AI use generally falls within professional discretion when it involves routine applications with appropriate safeguards and oversight. Explicit consent becomes appropriate when AI use is novel or experimental, when autonomous workflows might affect costs or timing significantly, when clients have expressed views about AI use, or when multi-step processes raise questions clients might reasonably want input on.

Beyond initial engagement, ongoing communication about agentic systems may be appropriate when autonomous workflows affect significant strategic decisions or when findings generated through multi-step processes will inform important client decisions. This communication should focus on what was done, why it's reliable, what limitations exist, and what the findings mean for the client's situation.

Looking Ahead

This chapter examined the ethical considerations unique to agentic AI systems—how autonomy affects professional responsibility, what transparency obligations arise, how accountability operates when agents make mistakes, what quality standards apply, where autonomy limits should be set, what competence requires, how to preserve professional independence, and what clients should know and consent to.

These ethical considerations are fundamental to responsible agentic AI deployment. Technology that violates professional obligations or compromises client interests fails regardless of its capabilities. The frameworks discussed in this chapter provide guidance for deploying agentic systems in ways that enhance legal practice while maintaining professional and ethical standards.



But ethical deployment is only part of ensuring these powerful systems serve client interests and advance justice. Part Three shifts focus to practical applications—how discovery professionals can actually use agentic and non-agentic AI to transform document analysis, transcript review, investigation workflows, contract management, and other core legal processes. The chapters ahead demonstrate how the technical capabilities from Part Two and the ethical framework from this chapter combine to enable new approaches to fundamental legal work.



PART THREE

PRACTICAL APPLICATIONS IN DISCOVERY AND INVESTIGATIONS

PART THREE

Practical Applications in Discovery and Investigations

Parts One and Two equipped you with a comprehensive understanding of Generative AI technologies—from foundational LLM concepts through advanced agentic architectures and reasoning models, from technical capabilities through ethical frameworks. Now we turn to the question that matters most for discovery professionals: How do you actually deploy these technologies to transform your work?

Part Three demonstrates practical applications across the discovery lifecycle. Rather than theoretical discussions of what AI might do someday, these chapters show what AI systems are doing now in real discovery workflows. The examples presented here are based on actual implementations, though details have been generalized to protect confidentiality and distilled to highlight patterns applicable across diverse matters.

For discovery professionals, this is where abstract capabilities become concrete. You'll see how workflows that traditionally require weeks of attorney time can be compressed to days, how synthesis that once demanded tedious manual work across dozens of depositions can now happen systematically, and how investigations transform when AI systems can autonomously pursue leads and identify patterns across massive document populations.

What You'll Learn

Part Three explores AI applications across the discovery lifecycle, from initial investigations and early case assessment through document review, transcript and medical records analysis, timeline construction, and trial argument preparation. Each chapter demonstrates practical deployment of the technologies introduced in Parts One and Two, showing detailed workflows, specific methodologies, actual results, and critical analysis of both strengths and limitations. The final chapter addresses the essential question of implementation: how to build AI-augmented discovery capabilities within your organization?

Why These Applications Matter

Understanding that agentic systems can plan and execute multi-step workflows is valuable. Knowing specifically how to deploy such systems to conduct investigations, analyze opposing counsel's production, build comprehensive timelines from scattered communications, or prepare closing arguments in minutes—that's transformative.

Successful AI deployment doesn't mean eliminating human involvement. It means augmenting human capabilities to deliver better outcomes more efficiently:

- Better early case assessment from rapid comprehensive analysis
- More thorough investigations as AI systematically identifies patterns manual review might miss
- Enhanced quality control through systematic verification

- Improved cost-effectiveness by focusing expensive human expertise where it adds most value
- Strategic advantages from comprehensive understanding of evidence enabling analyses that would be impractical with traditional methods

These outcomes require understanding not just what AI can do but how to deploy it effectively within the constraints of legal practice—which is precisely what Part Three provides.

How to Approach Part Three

As you read, consider your own matters and workflows. Where do you encounter document volumes that make comprehensive human review impractical? What investigations require following threads through multiple sources? What synthesis challenges demand coordinating information across dozens of documents? These are the applications where AI capabilities shine—and where Part Three’s insights will prove most valuable.

Pay particular attention to the verification protocols in each chapter. These aren’t optional extras—they’re essential components of responsible AI deployment in legal contexts. Every application requires human oversight, verification, and ultimate decision-making authority. The goal is augmentation, not automation.

Looking Ahead

The chapters ahead show how the architectures and capabilities from Parts One and Two manifest in concrete discovery applications. Part Three bridges the gap between capability and application, between understanding what agentic systems can do in theory and knowing how to deploy them effectively in practice.

The AI-augmented future of discovery isn’t coming—it’s here. These chapters show you how to be part of it.

Let’s begin where most matters begin: with investigations and early case assessment that shape strategic direction for everything that follows.

10

Using GenAI for Investigations and Early Case

CHAPTER TEN

Using GenAI for Investigations and Early Case Assessment

The Investigation Moment

The call comes at 4:47 PM on a Friday. Your company's general counsel needs answers by Monday morning. A whistleblower complaint alleges that sales executives discussed pricing with competitors at industry conferences. The board wants to know: Did this happen? How serious is it? Who was involved? Should we self-report to the DOJ? What's our exposure?

Or it's a different Friday afternoon crisis: a regulatory subpoena, a shareholder derivative demand, an internal audit red flag, a departing employee's concerning exit interview. The specifics vary, but the fundamental challenge remains the same. Critical questions demand immediate answers. Strategic decisions can't wait for traditional analysis. You need to know what happened, who did it, and what it means—now, not in ten weeks.

Traditional investigation faces predictable obstacles. Document volumes overwhelm manual review—fifty thousand emails and documents are a modest collection these days. Time pressure makes thorough analysis nearly impossible when racing toward board meetings, regulatory deadlines, or litigation decisions. Building keyword searches takes days of iteration, and even careful search development produces thousands of false positives requiring manual review. Deploying contract review teams for preliminary investigation costs \$30,000 to \$75,000 before you've made a single strategic decision. Most frustrating of all: you often can't define the scope until you understand the facts, but you can't understand the facts without reviewing the documents.

This creates an impossible equation. You need comprehensive answers to make informed decisions, but comprehensive analysis takes too long and costs too much. So you make preliminary judgments based on incomplete information, hoping you haven't missed something critical.

Using a RAG (or REG) system for early case assessment

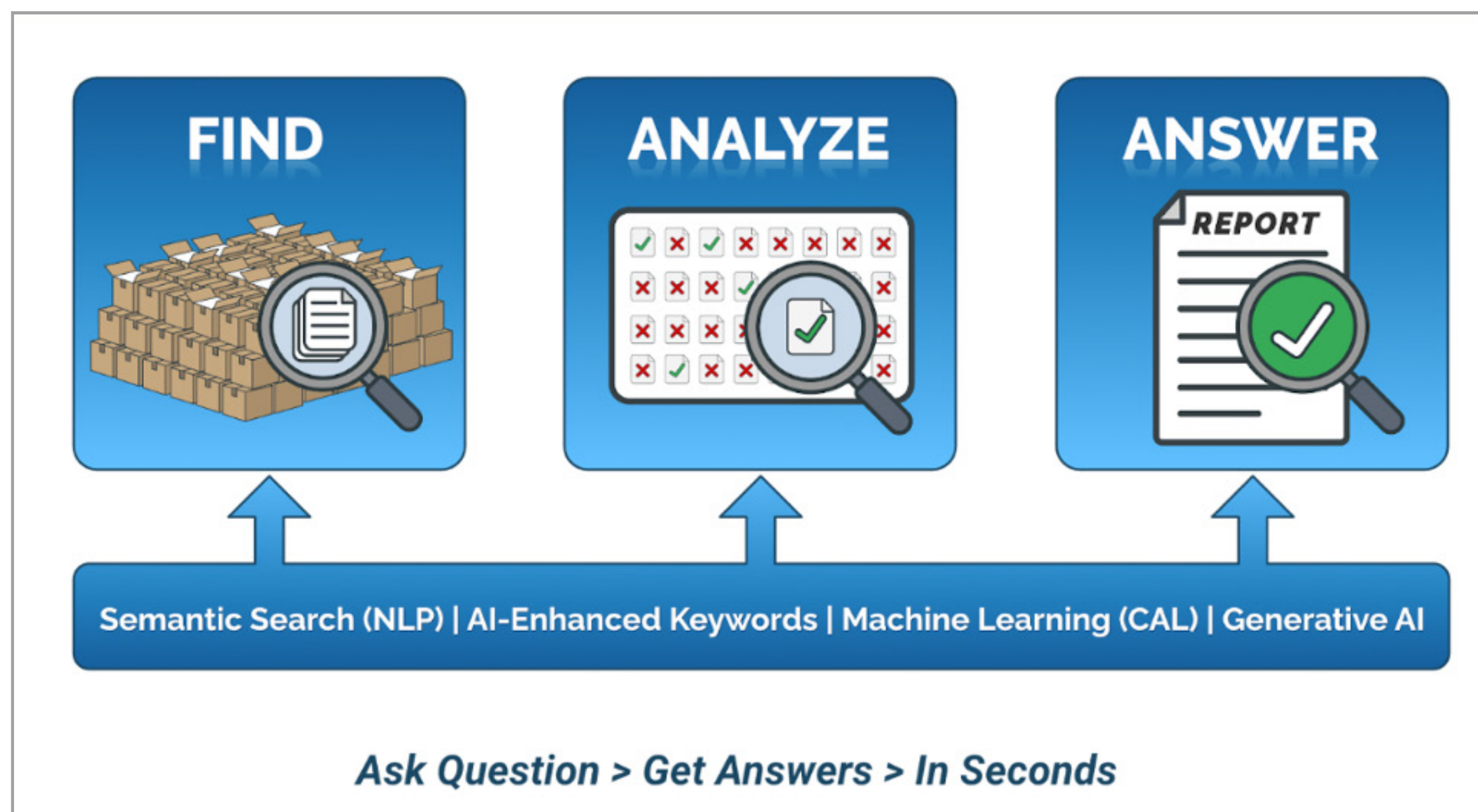
Generative AI transforms this equation completely. As discussed in Chapter 3, Retrieval-Augmented Generation (RAG) and Retrieval-Exclusive Generation (REG) systems combine the power of large language models with targeted search through your document collection, grounding AI analysis in actual evidence. What once took weeks of manual review now happens in hours of intelligent analysis. Instead of building complex keyword searches and reviewing thousands of false positives, legal teams ask questions in natural language and receive comprehensive answers with specific citations to supporting documents.

To illustrate how REG-based systems handle investigations in practice, we'll use Merlin Alchemy throughout this chapter. Alchemy's unique Find-Analyze-Answer framework demonstrates how modern AI platforms translate investigation questions into systematic document analysis. While specific implementations vary across platforms, the underlying approach—retriev-

ing relevant content, analyzing it with AI, and generating evidenced answers—represents how REG (and RAG) systems work in legal discovery contexts.

The techniques you'll learn here extend throughout the discovery lifecycle. Understanding how to frame investigation questions, interpret AI-generated answers, and verify conclusions with source documents supports not just initial assessment but document review decisions, deposition preparation, timeline construction, and trial argument development. But it all starts here, in those critical first hours when strategic decisions must be made and traditional methods simply take too long.

The Find-Analyze-Answer Framework



Alchemy operates through a three-stage framework that reflects how RAG systems approach investigation tasks: Find the relevant documents, Analyze their content systematically, and Answer your investigation questions with supporting evidence. Understanding each stage helps you use these tools effectively and interpret results appropriately.

Find: Three Types of AI-Powered Search

The Find stage employs three complementary search methods that work together to surface relevant documents without requiring complex Boolean queries or extensive keyword testing.

Semantic Search understands meaning and context, not just keyword matching. When you search for “pricing strategy,” the system finds documents discussing “market positioning,” “competitive rates,” “value proposition,” and “pricing approach” even when they don’t contain your exact search terms. The system recognizes that “Let’s be rational about Q4 pricing” expresses a concept relevant to price-fixing investigation even though it contains none of the typical antitrust vocabulary.

Consider a real example from the BP Oil Spill litigation. A lawyer posed this natural language question: “I need to analyze all technical testimony regarding the well flow indicators that were present before the Macondo well blowout, including data from monitoring systems, observed anomalies, and then identify all explanations given for why these signals were not adequately recognized or acted upon.” In seconds, the system’s three search types worked together seamlessly, finding the two hundred most relevant transcript sections. No complex Boolean query required. No days of keyword refinement.

AI-Enhanced Keywords expand search terms intelligently based on language patterns and contextual usage. The system identifies related terms and variations automatically. A search for “meeting” expands to include “call,” “discussion,” “conference,” “briefing,” “catch-up,” and other synonyms. More importantly, it adapts to how language is actually used in your document collection. If people write “sync” instead of “meeting,” the system learns this pattern.

Continuous Active Learning improves search effectiveness through feedback. As you review documents and mark them relevant or non-relevant, the system learns patterns that distinguish what matters from what doesn’t. After reviewing fifty documents, the system can re-rank the entire fifty-thousand-document population based on similarity to the relevant examples. The learning continues throughout investigation—each relevance judgment makes subsequent results better.

These three methods work together. Semantic search casts an intelligent net based on conceptual meaning. AI-enhanced keywords ensure comprehensive coverage across terminology variations. Continuous Active Learning refines results based on what you’re actually finding. The result: two hundred to five hundred highly relevant documents surface quickly from populations of tens of thousands, without complex Boolean queries or weeks of keyword testing.

Analyze: Systematic Document Processing

Once the Find stage retrieves relevant documents, the Analyze stage processes them through a two-level architecture that balances cost efficiency with analytical power.

The Analyze stage uses lightweight language models for high-volume document processing. Models like Claude Haiku or Gemini 2.5 Flash Mini read each retrieved document and generate structured summaries. They extract key facts—names, dates, dollar amounts, agreements, commitments. They score each document’s relevance to your investigation question on a numerical scale. They flag documents that appear privileged or confidential. They identify “hot documents” that warrant immediate attention based on content suggesting serious exposure or critical evidence.

This lightweight processing is cost-effective because these smaller models are twenty to fifty times cheaper than advanced reasoning models. A typical investigation might analyze five hundred documents for about \$11.00, a trivial expense compared to attorney review at hundreds of dollars per hour.

The result of the Analyze stage is a structured dataset: individual document summaries with extracted facts, relevance scores, risk flags, and categorizations. This structured data becomes the foundation for synthesis and strategic analysis in the Answer stage.

Answer: Synthesis and Comprehensive Reports

The Answer stage produces investigation findings in formats attorneys can use immediately for strategic decision-making. This is where advanced reasoning models provide synthesis and strategic analysis.

For synthesis, we recommend higher level reasoning models like Claude Sonnet 4.5, Gemini 2.5 Pro or GPT-5 to review the document summaries provided from the Analyze stage. They synthesize findings across the entire document set, identifying patterns, relationships, and contradictions. They draw evidence-based conclusions and assess exposure. They generate comprehensive investigation reports that organize findings thematically, cite specific supporting documents, and identify gaps requiring additional investigation.

This two-tier GenAI architecture makes AI-powered investigation economically viable. The lesser models handle the high-volume document processing efficiently. The higher models provide sophisticated analytical reasoning where it matters most—in synthesis and strategic assessment.

Reports begin with executive summaries that answer your investigation questions directly. Did employees discuss pricing with competitors? Yes—eight communications show senior executives discussing market price stabilization. Are these discussions defensible business intelligence or serious exposure? Serious—the language suggests agreement about future pricing behavior, not just market analysis.

Detailed findings follow, organized by theme or issue. Documents are categorized by risk level—red flags requiring immediate attention, yellow flags needing follow-up, green for defensible communications. Each category includes specific citations to supporting documents with brief summaries explaining why they matter. The system identifies key individuals involved and their roles. It constructs timelines showing when concerning communications occurred and whether patterns correlate with business events.

Prompt Helper: AI-Assisted Question Refinement

An important development in AI investigation platforms is systems that help you formulate better investigation questions. These prompt helper capabilities analyze your document collection, ask clarifying questions about your investigation objectives, and suggest search strategies based on available evidence. The result is more comprehensive investigation with less risk of missing critical information.

The process works through rapid document analysis and iterative refinement. You state your investigation objective in natural language—often just a sentence or two about what you need to know. The system performs a preliminary scan of the document population, identifying key themes, participants, time periods, and terminology patterns. Based on this quick assessment, it asks clarifying questions that help focus the investigation appropriately.

These questions target critical scope decisions: What time period matters most? Which individuals are key to understanding what happened? Are there specific transactions or events to focus on? What level of detail do you need? Consider a practical example. You receive a subpoena about potential FCPA violations related to South American operations. Your initial

question might be straightforward: “What evidence exists about improper payments to government officials?”

The system’s preliminary scan identifies references to consultants and third-party intermediaries in multiple countries, communications in English, Spanish, and Portuguese, and terminology like “facilitation fees,” “expediting payments,” and “relationship building” that might be euphemisms. It asks:

- Should the investigation include payments to intermediaries who may have government contacts?
- Should we analyze the full document time range or prioritize specific years?
- Should these euphemistic terms be included in the search?
- Should summaries flag original language for cultural context?

Your answers shape the analysis strategy:

- Include intermediaries? Yes
- Focus on 2020-2022 but flag patterns from earlier years? Yes
- Include euphemistic language? Definitely
- Flag original language? Good idea for avoiding misinterpretation

Based on these refinements, the system designs a comprehensive search strategy covering direct references to government officials, third-party consultant agreements, communications about compliance, internal audit reviews, and exchanges using identified euphemistic language. The result: four hundred to six hundred potentially relevant documents retrieved from fifty thousand in the collection, analyzed systematically, with findings organized to answer your investigation questions.

The preliminary questions took fifteen minutes. The comprehensive analysis completed in hours. The alternative—developing search terms manually, testing them iteratively, and potentially missing critical communications—would have taken days or weeks.

Investigation in Action: The Pricing Case

The Scenario

Your company received a Civil Investigative Demand from the DOJ Antitrust Division investigating potential price-fixing in your industry. The CID requests all documents related to pricing discussions with competitors. You have ninety days to respond. The document population includes seventy-five thousand emails and documents from eight custodians covering three years of communications.

The critical questions requiring immediate answers: Did any employee discuss pricing with competitors? If so, what was discussed and how serious is it? Who was involved and at what level of the organization? Is there evidence of competitive intelligence gathering versus actual coordination about future pricing? What’s the exposure? Should you self-report or pursue a voluntary disclosure strategy?

Traditional investigation would consume ten weeks and cost \$40,000 to \$80,000 in document

review before reaching strategic decisions. Custodian interviews, legal holds, keyword development and testing, contract attorney review of thousands of search hits, senior attorney analysis, and report preparation all take time.

AI-Powered Analysis

Upload the seventy-five thousand documents to the secure platform. In minutes, preliminary analysis identified three hundred fifty documents mentioning competitor names and twenty-eight specifically discussing pricing. Multiple communications involve trade association meetings and conferences where competitors were present.

You pose the investigation question Saturday morning: “Analyze all communications with competitors for evidence of pricing coordination and assess exposure level.” The system refines the question through brief clarifying exchange—should analysis focus on specific competitors named in the CID or all competitors? Should trade association contacts be treated differently from individual communications? What about indirect references to pricing through terms like “market discipline” or “rational pricing”?

Within 36 hours, comprehensive analysis is complete. The system identifies **eight documents** showing senior executives discussing “stabilizing market pricing” and “holding the line on 2023 pricing” in direct communications with competitor executives. These conversations occurred outside formal business contexts—after trade association dinners, through text message threads, in individual email exchanges. The participants included your VP of Sales and VP of Marketing communicating directly with their counterparts at competing firms.

Twenty-three additional documents require follow-up analysis—yellow flags showing detailed competitive intelligence shared at trade association meetings, social media exchanges about “market discipline,” and emails discussing competitors’ recent pricing moves with commentary that could be characterized as signaling.

The analysis constructs a **timeline** showing increased frequency of competitor communications in the first half of 2022, correlation with market price increases in the second quarter, compliance training conducted in the third quarter, and reduced competitor contact after the training. This pattern aligns precisely with DOJ’s investigation focus.

Key individuals are identified with their level of involvement. The VP of Sales participated in six of the eight red flag communications. The VP of Marketing attended trade association events where detailed pricing information was discussed. The CEO was copied on several yellow flag emails, raising questions about management awareness.

Exposure assessment: The red flag communications provide evidence DOJ could characterize as price-fixing conspiracy. While most communications are defensible business intelligence or legitimate trade association participation, the eight problematic documents present significant risk. The timing correlation with market events strengthens the government’s potential case.

The Strategic Advantage

This comprehensive intelligence enables immediate strategic decisions. You walk into Monday morning’s emergency board meeting with evidence-based answers, not preliminary guesses.



You can conduct privileged attorney interviews with the involved executives armed with specific document citations. Outside antitrust counsel can review the eight critical documents and assess voluntary disclosure strategy while timing still matters for cooperation credit.

The contrast with traditional investigation is stark: a weekend of AI-powered analysis versus ten weeks of preliminary review, minimal cost versus \$40,000 to \$80,000 spent before making strategic decisions, evidence-based exposure assessment versus educated guessing, proactive strategy development versus reactive scrambling.

The Continuous Intelligence Model

Traditional discovery treated each phase as separate and sequential: initial investigation led to decisions, then document review produced documents, then deposition preparation began from scratch, then trial preparation started over.

AI-powered investigation creates a fundamentally different model: continuous intelligence. The same system and data support progressively deeper analysis throughout the case life-cycle. Initial investigation findings inform document review decisions. Document analysis insights guide deposition preparation. Testimony analysis feeds trial argument development. Everything builds on what came before rather than starting fresh at each phase.

This continuity creates strategic advantages that compound over time. Patterns emerge across multiple analyses that wouldn't surface in isolated review phases. Contradictions between documents, testimony, and party positions become obvious when the system maintains comprehensive case knowledge. Each investigation question refines overall understanding. The comprehensive case knowledge that develops enables better decisions at every stage—from early exposure assessment through trial strategy and settlement evaluation.

Modern investigations involve diverse document types beyond email. AI platforms handle all text-based content: standard office documents (Word, PDF, Excel, PowerPoint) including comments and tracked changes; communication platforms (Teams, Slack, text messages) with conversation threading reconstructed automatically; transcripts from depositions, interviews, and meetings; and specialized documents like medical records, financial statements, technical reports, and regulatory filings. The system processes multiple languages, mixed file types, and handles conversion automatically. If you can read it, the system can analyze it.

The next chapters demonstrate how this continuous intelligence model transforms each subsequent discovery phase. Initial investigation capabilities extend to full document review workflows, testimony analysis across dozens of depositions, timeline construction, relationship mapping, and ultimately trial preparation where comprehensive record understanding delivers decisive advantage. But it all starts here, with investigation and early case assessment that establishes the foundation for everything that follows.

From Weeks to Hours, From Guessing to Knowing

Traditional investigation meant accepting painful constraints: waiting weeks for preliminary answers while strategic windows closed, spending tens of thousands of dollars before making strategic decisions, operating with incomplete information because thorough investigation cost too much or took too long, and making critical decisions based on educated guesses rath-

er than comprehensive evidence.

Those constraints no longer apply. AI-powered investigation delivers comprehensive analysis in hours, enables strategic decisions with full evidentiary support within days, and makes thorough investigation economically viable for matters that previously couldn't justify the cost. The analysis is often more thorough than manual review could achieve—systematic coverage of entire document populations, pattern recognition across hundreds of documents, consistent analysis without fatigue or oversight.

Integration with Professional Judgment

AI provides comprehensive analysis with citations but doesn't make strategic decisions—you do. It identifies potentially privileged documents, but edge cases require attorney judgment. It analyzes testimony for contradictions, but witness credibility demands human assessment. It reveals communication patterns, but organizational context requires your knowledge of the players and relationships. The most effective investigation approach combines AI-powered document analysis with traditional attorney skills. AI handles comprehensive document intelligence—reading, categorizing, analyzing, and synthesizing findings across large populations.

Senior attorneys focus on what AI cannot do: developing investigation strategy, conducting attorney-client interviews, making judgment calls on privilege and exposure, evaluating credibility and reliability of evidence, and making strategic decisions about voluntary disclosure, cooperation, or litigation approach.

The Competitive Reality

Leading law firms and corporate legal departments deploy these tools for investigations and early case assessment today. Government agencies are adopting AI-powered investigation capabilities. Your opposing counsel may already be using AI to analyze documents in cases against you, developing strategic advantage while you're still running keyword searches.

When that call comes—the regulatory inquiry, the whistleblower complaint, the litigation threat, the audit finding that demands immediate investigation—you face a choice:

Traditional Path: Ten weeks to preliminary findings. \$40,000 to \$80,000 in review costs before strategic decisions. Risk of missed critical evidence in time-pressured manual review. Decisions made with incomplete information. Reactive strategy development.

AI-Powered Path: 24-48 hours to comprehensive analysis. \$500 to \$2,000 in processing costs. Systematic coverage of entire document population. Decisions made with full evidentiary support and specific citations. Proactive strategy from day one based on comprehensive understanding.

The tools exist today. The platforms are proven in hundreds of investigations across major law firms, corporate legal departments, and government agencies. The capability isn't experimental—it's operational and delivering results. The question isn't whether AI will transform investigation and early case assessment. The question is whether you'll be among those leading the transformation or those scrambling to catch up.

Looking Ahead

This chapter focused on investigations and early case assessment—where you need answers fast and strategic decisions can't wait for traditional analysis timelines. The Find-Analyze-Answer framework that enables rapid investigation doesn't stop there. The same capabilities that help you understand a case in hours rather than weeks transform how you handle the most resource-intensive phase of discovery: document review for production.

Chapter 11 shows how GenAI changes document review from a brute-force reading exercise into an intelligent analysis workflow. The same techniques you use to investigate issues—asking questions, getting evidenced answers, verifying conclusions—apply to review decisions about responsiveness, privilege, and relevance. But the scale changes dramatically, the economic benefits multiply, and the quality improvements become essential to competitive practice.

11

Using GenAI for Document Review and Analysis

CHAPTER ELEVEN

Transcript and Testimony Analysis

Chapter 10 demonstrated how AI transforms investigation and early case assessment. That chapter showed how comprehensive intelligence can be delivered in hours rather than weeks. That rapid analysis enables critical strategic decisions when timing matters most. But investigation represents just the beginning. Once you understand what happened and assess exposure, you face the larger challenge: systematically reviewing and producing documents in response to discovery requests.

Generative AI is transforming how document review operates. Instead of armies of reviewers reading line by line, AI systems can analyze millions of documents in hours, identify responsiveness and privilege, summarize complex content, and apply consistent reasoning across every record. The result is faster, more accurate, and far more affordable review without compromising defensibility or quality.

This chapter explores how discovery teams are achieving these results today—and how AI is redefining what effective, intelligent document review looks like.

Document Review: The Cost Challenge

Document review has always consumed the majority of discovery budgets—roughly two-thirds of every dollar spent. On a typical matter, review costs range between one and one-and-a-half dollars per document, meaning that a one-million-document project can easily exceed \$1.25 to \$1.5 million and take three months to complete.

The reasons are straightforward: massive document volumes, labor-intensive processes requiring multiple judgments per document, high-quality standards with little margin for error, and complexity spanning languages, jurisdictions, and regulatory regimes. Traditional review scales linearly with document count—double the documents, double the cost and time.

Generative AI changes that calculus completely. Modern review platforms can analyze and classify documents at ten to twenty-five cents per document and compress timelines from months to weeks. For a million-document matter, costs can drop to \$100,000 to \$200,000, with completion in three weeks rather than three months. Unlike earlier discovery tools that could only search and filter, Generative AI reads documents for meaning and context, enabling it to apply uniform standards across entire populations and turn scale from an obstacle into an advantage.

Using GenAI to Power Production Review

Production review represents the core function of document review in discovery. When legal teams face document requests from opposing counsel or government agencies, they must identify responsive documents, protect privileged materials, redact confidential information, and tag documents by issue. Getting this wrong has consequences ranging from discovery

sanctions to inadvertent privilege waivers to producing documents that damage case strategy. Traditionally, this work required armies of contract attorneys reviewing documents one by one, making coding decisions based on detailed review protocols. A significant matter might involve hundreds of reviewers working for months, generating millions in review costs. Even with this investment, consistency problems plague large review teams as different reviewers apply varying judgment to similar documents.

Generative AI transforms production review by bringing consistency, speed, and analytical depth that human reviewers working alone cannot match. AI can analyze documents in seconds rather than minutes, apply uniform criteria across massive populations, and flag complex issues that reviewers might miss. The technology doesn't replace human judgment, but it dramatically reduces the volume requiring detailed human review while improving quality. This section examines how AI enhances the four critical components of production review: responsive review, privilege review, redaction review, and issue review.

Responsive Review

Responsive review answers a fundamental question: does this document respond to the discovery request? This determination drives everything else. Non-responsive documents need no privilege review, no redaction, no issue coding. Getting responsiveness wrong wastes resources on unnecessary review or, worse, fails to produce required documents.

Discovery requests vary in breadth. Some requests seek narrow categories like "all communications between Person A and Person B regarding a specific transaction." Others cast wider nets like "all documents concerning, relating to, or mentioning the merger." And some requests are multi-part, covering tens and even scores of different or partially-related topics. In each case, reviewers (and their managers) must understand the request scope, interpret document content, and make consistent judgments across thousands or millions of documents.

How AI Performs Responsive Review

AI-powered responsive review uses LLMs to read documents and make determinations about responsiveness, privilege, and relevance. Unlike traditional TAR systems that rely on statistical patterns, GenAI actually reads document content and evaluates it against discovery request criteria in much the same way a contract attorney would.

The workflow starts with careful translation of discovery requests into evaluation criteria. Rather than relying on keyword searches that produce both relevant documents and massive false positive populations, AI systems read documents for meaning and context. They understand nuance, recognize relevant concepts expressed in different ways, and evaluate documents based on substantive content rather than mere word matching.

The system receives discovery requests as prompts defining what makes documents responsive. For a request seeking "communications regarding pricing strategy for Product X," the prompt instructs the AI to identify documents discussing pricing decisions, competitive pricing analysis, price-setting methodology, or pricing discussions specifically about Product X. The prompt includes examples of clearly responsive, clearly non-responsive, and borderline documents to calibrate the system's judgment.

Documents flow through this evaluation systematically. The AI reads each document, assesses

its content against the responsiveness criteria, and generates a determination with reasoning.

A responsive document receives a “Responsive” tag and an explanation of why it meets the criteria. A non-responsive document receives a “Non-Responsive” tag with reasoning. Documents where responsiveness is unclear receive an “Uncertain” tag for human review.

This approach delivers several advantages over traditional review. Speed increases dramatically—AI analyzes documents in seconds rather than the minutes required for human review. Consistency improves because the same criteria apply uniformly across all documents rather than varying based on which reviewer sees which document. Coverage becomes comprehensive because AI doesn’t suffer attention lapses or fatigue that cause human reviewers to miss responsive content.

Organizing AI-Powered Responsive Review

There are two principal ways to deploy Generative AI in production review, depending on how much human oversight is built into the workflow.

The first approach keeps a human reviewer in the loop for every document. As each document appears on screen, the AI engine reads it instantly, applies the review criteria, and presents a recommendation—responsive, non-responsive, or uncertain—along with a concise explanation of its reasoning. The reviewer can accept the recommendation, override it, or flag the document for further analysis. This model ensures that human eyes see every document and that reviewers remain fully engaged in decision-making. However, it also slows the pace of review and increases costs. Each reviewer must pause to read the AI’s recommendation and reasoning, and the additional computational step adds incremental API expense. For some matters, that trade-off is justified; for most, it is not.

The second approach—the one we recommend—places the AI engine in the primary review role. Here, the system, not individual reviewers, analyzes the document population and makes the responsiveness determinations directly. Human expertise shifts upstream and downstream: review managers design and test the prompts, calibrate confidence thresholds, and perform continuous quality control throughout the project. Human reviewers then perform targeted sampling and final QC testing to validate accuracy. This model preserves defensibility while realizing the full efficiency gains of Generative AI. Review teams focus on oversight, verification, and refinement rather than repetitive document-by-document reading.

We recommend this model for several reasons. First, it is faster—AI can analyze documents at a rate no human team can match, reducing review timelines from months to weeks or even days. Second, it is more accurate and consistent. Our research and production experience show that top large language models match or exceed the accuracy of skilled legal reviewers while applying standards uniformly across every document. Third, it is dramatically more cost-effective. Removing the need for human reviewers on each document eliminates the largest cost component of traditional discovery while still maintaining defensibility through structured quality control.

Whichever model is used, effective responsive review depends on careful organization and workflow design. Several key decisions determine how the process operates.

Organizing the Workflow

Once the review model is defined, success depends on structure. Large reviews move fastest when they follow an organized, auditable process. Teams begin by prioritizing documents likely to hold the greatest value—those from key custodians, critical time periods, or containing sensitive terms—so that early results deliver maximum insight. The collection is divided into manageable batches that allow steady progress and continuous quality checks rather than overwhelming the system with millions of files at once.

Confidence thresholds guide which documents flow automatically through the process and which require human verification. High-confidence, non-responsive materials can be set aside with minimal oversight, while documents that trigger uncertainty receive additional review. As results accumulate, prompts are refined and thresholds adjusted, steadily improving accuracy.

Verification remains essential. Early batches are sampled more heavily to establish baseline performance; once the system proves reliable, lighter ongoing sampling maintains oversight without slowing progress. The goal is a process that is fast, consistent, and effective—one that blends automation with the discipline of continuous human quality control.

Optimizing AI-Powered Production Reviews with Continuous Active Learning

The most efficient large-scale reviews use Continuous Active Learning (CAL) to direct the work dynamically. As introduced in Chapter 10, CAL continually re-ranks the document population based on feedback from prior judgments. Instead of reviewing every document, each coding decision—whether made by an AI model or a human verifier—is immediately fed back into the CAL algorithm, which updates its ranking of the entire corpus in real time.

The effect is dramatic. Responsive documents quickly rise to the top of the queue, allowing reviewers to find the vast majority of relevant material after examining only a fraction of the population—often as little as ten percent. Because the system can re-rank a million documents in roughly a hundred milliseconds, it adapts almost instantly to each new decision, continuously refining its understanding of what matters and what does not.

This rapid feedback loop makes CAL-driven review highly efficient. Reviewers see the most relevant documents first, enabling early case assessment and faster decision-making while avoiding the cost of reading vast numbers of irrelevant files. When the system reaches the point where newly reviewed documents yield few or no additional responsive items, an elusion review is conducted to confirm that stopping is justified. This statistical sampling verifies that the remaining population contains no significant undiscovered responsive material, ensuring both efficiency and defensibility.

Other Types of Review

Beyond responsive review, AI provides substantial assistance for privilege review, redaction review, and issue coding. Each application follows similar principles: AI identifies documents or content requiring attention based on defined criteria, explains its reasoning, and enables human reviewers to focus on judgment calls rather than mechanical screening.

Privilege Review

This type of review represents document review’s highest-stakes component. Producing privileged documents can waive attorney-client privilege or work product protection permanently. Over-designating documents as privileged wastes time defending privilege designations and may draw challenges from opposing counsel. The volume of documents requiring privilege assessment in modern litigation makes thoroughness difficult and expense prohibitive.

Traditional privilege review requires attorney reviewers trained in privilege law to examine every potentially privileged document. Even with training, consistency suffers as different reviewers apply varying judgment to similar facts. The pressure to review quickly creates risk that privileged documents slip through while time spent on clearly non-privileged documents wastes resources.

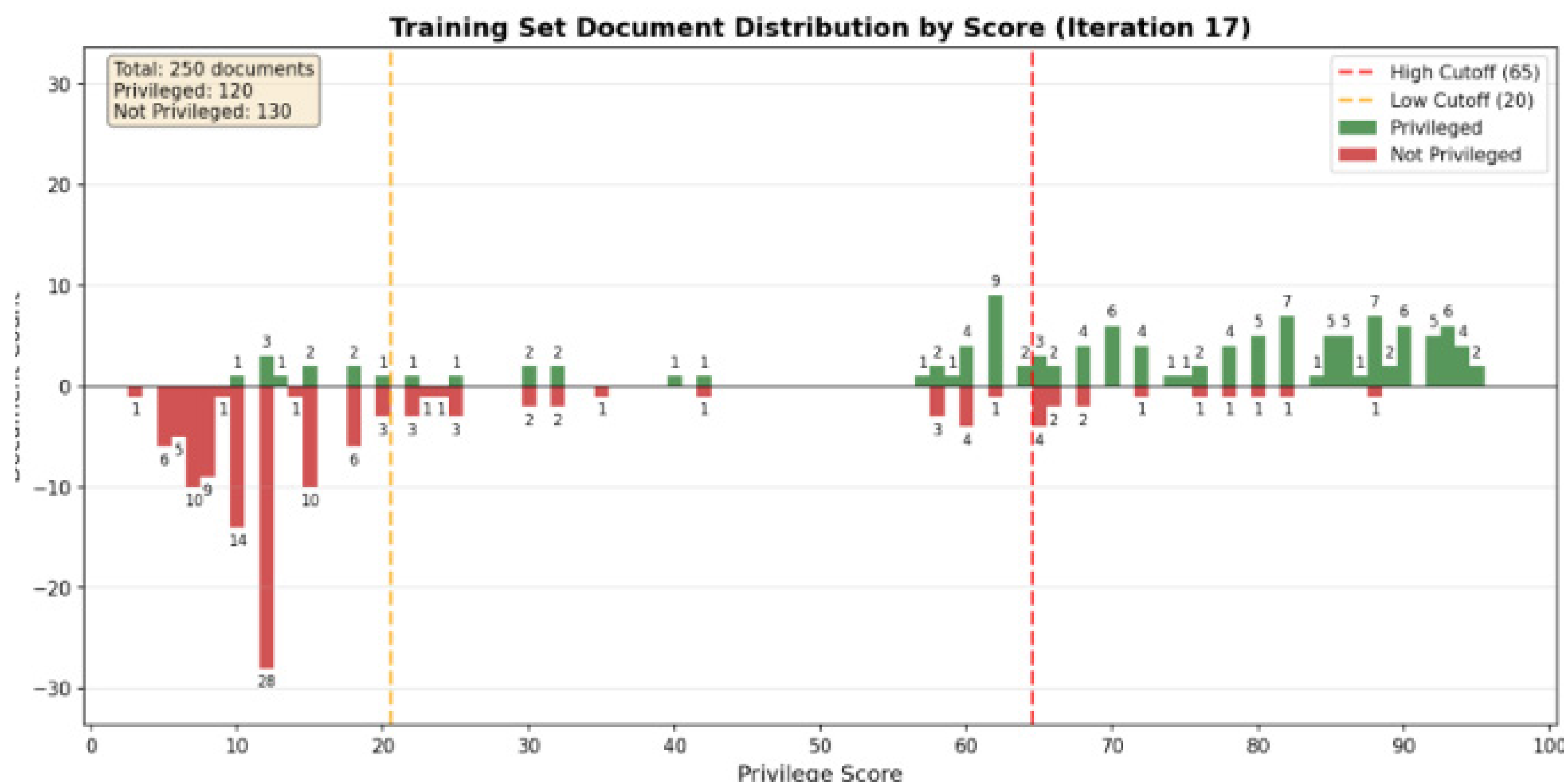
The critical question is whether AI can reliably identify privileged documents given the legal complexity and high consequences of errors. The answer is yes, but with important qualifications. Success requires using the most capable LLMs available, combined with carefully engineered prompts that capture privilege doctrine’s nuances.

Successful prompts require comprehensive context beyond keyword matching: attorney names and roles, law firm relationships, coverage counsel, litigation consultants, and the factual circumstances under which privilege attaches. The prompts should distinguish between legal advice and business advice, identify work product created in anticipation of litigation, and recognize when the crime-fraud exception or other privilege limitations may apply.

We use reasoning models to provide this level of analysis. Here is an anonymised report from a privilege review of a complicated set of business interactions just to give you an idea of the type of analysis top of the line GenAI reasoning models can offer.

Score	Decision	Rationale
25	Not Privileged	This is a credit committee agenda and meeting minutes for business credit decisions. While the Deputy General Counsel is copied on the email, the document is primarily operational/commercial in nature, discussing credit approvals, facility terms, and business strategy. No legal advice is being sought or provided, and the content focuses on credit risk assessment and business decisions rather than legal matters.
90	Privileged	This is a business email chain containing weekly call notes and operational updates between Acme executives. While it involves some Acme legal team members (Sean Jones, John Taylor), they appear to be receiving business information rather than providing legal advice. The content is purely commercial - funding updates, client meetings, deal progress - with no legal advice, analysis, or privileged communications present.
15	Not Privileged	This is primarily a commercial email chain discussing business terms for a financing facility with Acme Financial Corp. While it includes two identified Ace Corporation legal team members (Frank Russel and Joe Rollins), they appear to be involved in their business capacity rather than providing legal advice. The content focuses on commercial structuring, pricing, and business mechanics rather than legal advice or analysis.
85	Privileged	This email chain involves communications between Acme’s legal team (Joe Johnson - Deputy General Counsel, Robert Smith - General Counsel) and external law firm Dewey & Morgan (DMLaw.com) regarding legal advice on insurance policy documentation for a new investment. The dominant purpose is providing legal advice on insurance coverage requirements, with external counsel requesting specific legal documents to advise their client.

When implemented with this level of specificity and appropriate reasoning models, AI can reliably identify privilege indicators and flag potentially privileged documents for attorney review. In our experience, properly-trained LLMs match or exceed the performance of skilled legal reviewers while applying standards uniformly across every document. Well-orchestrated systems reliably identify clearly privileged and non-privileged documents, leaving a fraction of the review population for higher-cost attorneys to review and categorize the closer calls.



This tiered approach focuses expensive attorney time on genuinely ambiguous privilege determinations while quickly clearing obviously non-privileged business communications. AI can also draft privilege log entries by extracting document metadata and generating descriptions, which attorneys then verify and refine. The consistency advantage is significant: uniform privilege criteria applied across all documents reduce both the risk of missing privileged content and the inefficiency of over-flagging routine business communications.

Redaction Review

Reviews often include the need to identify specific text requiring redaction before production—trade secrets, personal information, financial data, or content covered by protective orders.

Traditional redaction review requires reviewers to read responsive documents, identify information requiring redaction, and ensure redactions are applied correctly and consistently. This is time-consuming and consistency is difficult to maintain. Redaction standards may be complex, reviewers interpret standards differently, and the volume of documents makes comprehensive review challenging.

AI engines can flag passages meeting redaction criteria and explain why each passage requires redaction. In many cases, human reviewers are used to verify flagged content, ensure completeness, and apply redactions. The primary benefit is consistency: all Social Security numbers are identified regardless of format, all revenue figures are flagged regardless of presentation. This uniform identification enables consistent redaction decisions across the production set.

Issue Coding

Moving beyond production, teams must organize responsive documents by case issues to support focused analysis, deposition preparation, and trial presentation. With a suitable prompt, an AI engine can evaluate documents against detailed issue definitions and assign relevant tags with explanations. Clear, specific issue definitions produce better results than vague categories. The system can handle hierarchical tagging (major categories with subcategories) and multi-issue documents. Well-organized issue tagging enables attorneys to quickly locate all documents relevant to specific issues, recognize patterns within issue categories, and assess evidence strength for settlement discussions.

For all three applications, verification protocols should be proportionate to risk. Privilege review demands higher sampling rates initially—perhaps 5-15% until accuracy is established—because errors carry severe consequences. Redaction and issue coding can use lighter initial sampling of 5-10%, adjusting based on observed accuracy. Track error rates by document type and category to identify areas needing enhanced scrutiny. As consistent accuracy is demonstrated, reduce ongoing sampling to 3-5% for continuous monitoring while maintaining higher rates for high-stakes matters or new document types.

Using GenAI for Multi-Language Review

Cross-border investigations and international litigation often involve documents in multiple languages—German, French, Mandarin, Portuguese, and beyond. Traditional multilingual review is both costly and cumbersome: teams must either hire legally trained reviewers fluent in each language or translate entire populations at significant expense, delay, and risk of inaccuracy.

Generative AI changes this equation. Modern large language models can read documents in their original languages, assess responsiveness and privilege, and summarize findings in the reviewer's preferred language—all without separate translation. Prompts written in English define relevance, privilege, and issue criteria; the system applies those standards to every document it can interpret and returns clear English analyses and summaries.

The benefits are immediate. Reviewers can evaluate communications and technical reports in dozens of languages without language fluency. Consistency improves because the same criteria apply across all languages, avoiding the variability that plagues multilingual human teams. The system can even suggest new issue tags when it detects topics not explicitly defined, helping uncover unexpected themes.

Verification remains essential. Performance varies across languages and specialized terminology, so review teams should test accuracy on representative samples. For less common languages or highly technical materials, human verification ensures reliability. Cultural context also matters—differences in tone, hierarchy, and communication style can affect interpretation—so quality control must account for those nuances.

Consider a cross-border antitrust investigation. The AI system reviews English, German, Mandarin, and Portuguese documents simultaneously, identifying communications about pricing or market coordination, summarizing them in English, and flagging potentially privileged content. Senior attorneys then review high-priority materials and translate only the few that war-

rant full examination.

What once required multiple language teams and months of translation can now be completed in weeks. Generative AI delivers a consistent and cost-effective process for multilingual review—turning what was once a logistical obstacle into a routine part of discovery.

Quality Assurance in AI-Assisted Document Review

No matter how advanced the technology, quality assurance remains essential. Every AI-assisted review must demonstrate that results are accurate, consistent, and defensible.

Verification begins early. Initial batches receive higher sampling to confirm the model's accuracy—typically ten to fifteen percent across responsive, non-responsive, and uncertain categories. Once accuracy stabilizes, sampling can drop to around five percent for continuous monitoring. For new document types, high-stakes issues, or unfamiliar languages, higher verification rates are appropriate until confidence is established.

Quality control focuses on pattern recognition rather than exhaustive checking. Managers track error rates by custodian, document type, and confidence level to identify where prompts or thresholds need refinement. When discrepancies arise, they analyze whether the issue lies with the model, the data, or the instructions and adjust accordingly.

Continuous monitoring sustains reliability over time. Periodic sampling and disagreement analysis between AI and human reviewers help detect drift or bias. The goal is not to re-review every document but to confirm, through consistent evidence, that the system is performing as intended.

Done properly, quality assurance converts automation into assurance—verifying that speed and efficiency never come at the cost of accuracy or defensibility.

Looking Ahead

The transformation of document review is already well underway. What began as experimental automation has matured into a proven, defensible workflow—faster, more consistent, and dramatically more cost-effective than traditional methods. For discovery professionals, the challenge now is not whether to adopt AI but how best to integrate it into everyday practice.

This chapter examined how Generative AI is transforming document review—reducing cost, accelerating timelines, and improving consistency while maintaining defensibility and quality. These capabilities have redefined what efficient, intelligent discovery looks like. Yet documents represent only part of the story. Testimony—depositions, hearings, and interviews—holds the narratives that shape cases. The next chapter turns to transcript analysis, showing how the same generative tools used to classify and summarize documents can extract key testimony, map issues, and connect facts across witnesses.

The evolution from document review to full-spectrum discovery intelligence is already in motion. Organizations that embrace these tools now will not only work faster but think faster, gaining insight and advantage long before their competitors.

12

Using GenAI for Transcripts and Medical Record Analysis



CHAPTER TWELVE

Using GenAI for Transcripts and Medical Record Analysis

The executive's deposition runs over 250 pages. The plaintiff's medical records span 12,000. Both contain critical evidence—testimony about what the executive knew and when, clinical documentation of causation and damages. Traditional analysis meant weeks of manual review by lawyers and nurses, creating summaries that became outdated as new depositions occurred. The costs to analyze and summarize the files could run to the tens of thousands.

Generative AI transforms narrative document analysis. What once required weeks of professional time now happens in hours. Instead of reading page by page, legal teams ask questions and receive comprehensive answers with specific citations. The system processes thousands of pages, extracts temporal patterns, identifies contradictions across witnesses, and maintains dynamic intelligence that updates as new testimony emerges.

This chapter examines how the Find-Analyze-Answer framework from Chapter 10 extends to testimony and clinical records—enabling discovery professionals to extract intelligence from narrative documents at scale while maintaining the citation accuracy and defensibility that litigation demands.

The Traditional Constraint

Traditional deposition and medical record review consumed enormous resources. Junior lawyers, paralegals, or nurse consultants read each page, creating summaries through manual note-taking and cross-referencing. The process was expensive, inconsistent, and slow. A paralegal might summarize one 200-page deposition in 6-8 hours. Scale that across dozens of witnesses and the timeline becomes measured in weeks.

Keyword searches helped locate specific terms but rarely delivered understanding. They couldn't explain context, recognize contradictions across witnesses, or connect cause and effect across multiple documents. True comprehension required human reading—until now. Modern GenAI platforms process thousands of pages in minutes, producing structured summaries with direct citations that make every analytical point instantly verifiable.

AI-Powered Transcript Summaries

When transcripts load into a GenAI-integrated discovery platform, the system divides each transcript into logical sections—by question sequence, subject, or topic break. This segmentation makes search efficient and ensures only relevant text flows to the language model for analysis.

The system then produces structured summaries, each linked directly to the source transcript. Every point includes a citation back to specific page and line numbers. Attorneys navigate from summary to source with a single click, maintaining the evidentiary chain of custody that litigation demands.

Example: Joseph Nadeau’s Deposition Testimony

Joseph Nadeau worked at Metro-Atlantic Chemical Company in 1964, mixing chemical ingredients for a plant later challenged for environmental pollution. The following examples show summaries produced entirely by an LLM with no human editing, demonstrating the analysis an LLM can produce in minutes at costs substantially below traditional legal professional rates.

Table of Contents

Here is an opening table of contents:

Deposition Summary Joseph Nadeau Case name: Emhart Industries, Inc. v. New England Container Company, Inc., et al. Date: Wednesday, June 12, 2013 Witness: Joseph Nadeau Location: offices of Duffy & Sweeney, Ltd, 1800 Financial Plaza, Providence, Rhode Island TABLE OF CONTENTS Overall Summary Timeline Key People Exhibits Table of Sections Transcript Sections Transcript Text

Page and Line Summaries

The system provides summaries of Nadeau’s testimony on dangerous incidents at the plant, with blue hyperlinks providing specific page and line references:

Nadeau describes multiple dangerous incidents at Metro-Atlantic, including a major methanol fire when a delivery driver ignited 3,000 gallons with a space heater. He rescued his brother from the liquid fire by shutting off the methanol valve, pushing the flames into French drains leading to the river. Other incidents included smaller fires, a chemical explosion that blew out building walls and required neighborhood repainting, and flooding that caused barrels to float downstream.	
101:11–101:24	Multiple fires occurred at Metro-Atlantic during Nadeau's time there, but he does not recall any fires at NECC
102:1–102:14	A major methanol fire occurred when a delivery driver put a lit space heater into a tank containing 3,000 gallons of the highly flammable methanol through an open valve
102:14–103:4	The methanol fire created a liquid fire 4-6 inches deep flowing out of the tank, with Nadeau's brother Bob trapped inside trying to fight the fire with a hose

103:5–103:20	Nadeau rescued his brother by using a hose to create a path to the valve to shut off the methanol flow, pushing the floating fire into the French drain where it flowed to the river and burned off
104:1–104:16	Other smaller fires occurred at the Metro building, including one that caused Nadeau to jump out a loading ramp window into a Dumpster to escape
104:17–105:7	An explosion occurred at Metro-Atlantic when someone mixed incompatible chemicals in a storage tank, blowing out a chunk of wall and requiring neighborhood houses to be repainted
105:1–105:7	The explosion blew off a submarine hatch-like cover 6-8 inches thick that was never found, snapping one-inch bolts and sucking out the side of the building
105:8–105:25	Flooding occurred when the river got high, causing barrels stacked behind Metro to float around the area and down the river, though Nadeau heard about this from coworkers rather than witnessing it directly
106:1–106:9	Nadeau recalls seeing the parking lot flooded during at least one flood, but it did not stop work operations or prevent access to the building where he worked

Unlike manual summaries, these links take the reader directly to the specific page of the transcript containing the summarized testimony.

Timelines and Multiple Views

The system can extract every temporal reference from a deposition and organize them chronologically. When a witness testifies about events spanning months or years, these references appear scattered throughout hundreds of pages. The AI identifies each mention of dates, times, or temporal relationships (“before the meeting,” “two weeks later,” “sometime in March”) and compiles them into a coherent timeline.

Timeline

1950

- Technical Manual and Yearbook of the American Association of Textile Chemists and Colorists published, later referenced as Exhibit 5 ([86:15–87:1](#))

1962-1963

- Nadeau begins working at New England Container Company (NECC) around 1962 or 1963 as a summer job, working one or two summers until school started ([16:23–17:8](#))
- Nadeau works part-time during summers of 1962 and 1963 at NECC ([27:16–27:21](#))

1964

- Nadeau graduates high school in 1964 ([13:7–13:25](#))
- Nadeau works part-time during summer of 1964 and becomes full-time in 1964 ([27:16–27:21](#))
- Nadeau works at Metro-Atlantic starting in 1964 after graduation, from June through August 1964 ([20:1–20:19](#))

August 23, 1965

- Nadeau is inducted into the U.S. Army, serving until August 22, 1967 ([15:9–15:24](#))
- Nadeau worked at Metro-Atlantic until August 23, 1965 ([107:24–108:1](#))

April 5, 1965

- Aerial photograph of the Centredale Manor site taken (Exhibit 7) ([133:14–134:6](#))

1965

- Sanborn map of the Centredale site dated 1965 (Exhibit 6) ([106:22–107:23](#))
- Nadeau becomes full-time employee in 1965 ([27:16–27:21](#))

This content has been generated by an artificial intelligence language model. While we strive for accuracy and quality, please note that the information provided may not be entirely error-free. We recommend independently verifying the content. We do not assume any responsibility or liability for the use or interpretation of this content.

The timeline captures not just dates but also the witness's certainty or uncertainty about timing. When testimony says "I think it was March, maybe April," that ambiguity is preserved. When testimony about the same event appears in multiple places with slightly different dates, those variations are flagged for investigation.

This timeline generation takes minutes with complete citations. A paralegal creating this same timeline manually would typically spend 6-8 hours reading, noting references, organizing chronologically, and formatting—work that now happens in minutes while maintaining citation accuracy.

People, Exhibits, and Topics

The system extracts and organizes testimony about specific individuals—who the witness mentioned, what they said about each person's role and actions, what meetings or communications involved particular people. The output is a person-by-person index with full context and citations. For a deposition mentioning dozens of people across hundreds of pages, this makes testimony about key players immediately accessible.

The system also tracks every exhibit—documents shown to the witness, discussed, or mentioned. Did the witness authenticate the document? Dispute it? Provide context about when it was created? Each exhibit gets a summary of all related testimony with links back to relevant sections.

Exhibits

Exhibit 1

Nadeau's deposition transcript from December 17, 2002 in Emhart Industries, Inc. vs. Home Insurance Company and various other insurance companies (10:16–11:7). Nadeau recalls providing this deposition testimony and confirms he was under oath and testified truthfully to the best of his knowledge at the time (11:8–11:23).

Exhibit 2

Nadeau's deposition transcript from November 24, 2008 in the case Emhart vs. New England Container (12:1–12:8). Nadeau recalls attending this deposition and confirms he was under oath and tried to provide truthful testimony to the best of his knowledge (12:9–12:20). Counsel later references pages 42-43 of this exhibit to refresh Nadeau's memory about stacking drums near the Texas Tower (126:3–126:19).

Exhibit 3

Document with Bates stamp E-000003.003225 through 3227 that Nadeau was shown in a prior Home Insurance deposition (22:22–23:19). This is an affidavit that Nadeau recognizes his signature on and recalls reviewing and signing under pain and penalty of perjury on January 20, 2001 (25:1–25:23). The document was written by an investigator who came to Nadeau's house and interviewed him about his time at NECC and Metro-Atlantic (26:22–27:5). The affidavit describes his employment from 1962-1965, his duties at New England Container including loading barrels into furnaces, and details about barrel reconditioning and waste disposal (27:8–40:20).

This content has been generated by an artificial intelligence language model. While we strive for accuracy and quality, please note that the information provided may not be entirely error-free. We recommend independently verifying the content. We do not assume any responsibility or liability for the use or interpretation of this content.

Beyond standard categories, the system organizes testimony by any topic relevant to the case—all testimony about the board meeting, everything the witness said about safety protocols, every reference to the contamination incident. These topic summaries capture testimony that might appear in response to questions about seemingly unrelated subjects, gathering scattered references into a single, coherent view.

Speed, Scale, and Comprehensiveness

A properly architected system using serverless computing can scale without limits. In one case, we created 102 deposition summaries in 131 minutes, with each running in separate computing space covering 18,000 pages of testimony.

102 Transcripts in 131 Minutes

A national law firm faced an enormous discovery challenge: 102 depositions totaling more than 18,000 pages. Traditionally, this task would have required weeks of attorney and paralegal time and cost hundreds of thousands of dollars.

Using a GenAI-integrated discovery platform, the firm transformed the process. Each transcript was automatically sectioned, summarized, and linked. In 131 minutes the platform produced complete summaries for all 102 depositions—structured outlines, topic tables, and page-line hyperlinks back to the original text.

For a small fraction of the cost of manual summaries, the firm obtained consistent, audit-ready summaries overnight. Traditional manual review of this volume would require approximately 400-600 attorney and paralegal hours at standard billing rates, generating \$80,000-\$150,000 in costs over 2-3 weeks. The GenAI approach reduced costs by 95% with overnight delivery. Reviewers could immediately search testimony across all witnesses and begin analysis the next day.

Richer Detail in Less Time

This multiplicity of views demonstrates a crucial advantage: AI summaries can be more comprehensive than human summaries. Traditional summarization involved painful trade-offs between detail and deliverability. GenAI removes this constraint. A summary covering every substantive exchange, every person mentioned, every temporal reference, and every exhibit discussion—organized simultaneously by chronology, topic, and person—can be generated overnight. The system doesn't tire, doesn't skip passages, and doesn't make judgment calls about what might matter later.

When attorneys need more detail about any section, they ask follow-up questions and receive expanded summaries with additional context, all still linked to source pages and lines. The depth of analysis becomes limited only by what questions the user asks, not by how much time a paralegal has available.

Medical Records: Parallel Architecture, Clinical Complexity

Medical records present the same narrative document challenge as depositions—enormous volumes of important evidence locked in narrative form. A single hospitalization can generate thousands of pages. Hospital records often dwarf transcripts, with some spanning 20,000 pages across multiple facilities and providers.

Unlike transcripts, medical records contain specialized clinical terminology often incomprehensible to legal professionals. They resist traditional keyword search because medical lan-

guage varies widely—the same condition might be described using formal diagnoses, clinical abbreviations, or colloquial shorthand. “Myocardial infarction,” “MI,” “heart attack,” and “acute coronary event” all describe the same medical reality but require different search terms.

Medical records are traditionally reviewed and summarized by medical professionals—nurses or nurse consultants—either onshore or offshore at high hourly rates. The costs for manual summarization typically exceed those for deposition summaries, and the output is often a simple chronology with limited analytical depth.

AI-Powered Medical Record Analysis

The same architecture that transforms testimony analysis applies directly to complex medical records. The system uses four rounds of GenAI analysis: first breaking records into individual “encounters” (doctor visits or medical procedures), then summarizing each encounter, then determining which are most important to the overall case, and finally creating the comprehensive report.

Modern LLMs are sophisticated enough to identify conditions, treatments, and causal relationships; compile patient timelines; extract medications and imaging results; and link every statement back to the original source document. Because the language model can interpret medical records as a clinician would—understanding lab results, imaging findings, and clinical diagnoses—it can also explain them in plain terms to lawyers, adjusters, or jurors who lack medical training.

Here is an example of a summary table of contents:

Deposition Summary Joseph Nadeau Case name: Emhart Industries, Inc. v. New England Container Company, Inc., et al. Date: Wednesday, June 12, 2013 Witness: Joseph Nadeau Location: offices of Duffy & Sweeney, Ltd, 1800 Financial Plaza, Providence, Rhode Island TABLE OF CONTENTS Overall Summary Timeline Key People Exhibits Table of Sections Transcript Sections Transcript Text
--

In this anonymized example, the LLM condensed 2,400 pages of hospital, billing, and consultation notes into a comprehensive, citation-linked report. The summary included an overview of the accident and resulting medical complications; key conditions such as respiratory failure, pulmonary hypertension, and diabetes; a timeline of treatment including imaging, medications, and consultations; key practitioners and facilities with their roles; and noted issues of

potential liability such as medication delays and discharge planning.

The system produces detailed narrative timelines far more comprehensive than typical manual summaries:

Timeline of Conditions and Treatments

September 2022

- **09/10/2022:** The patient was involved in a motor vehicle accident, which led to her subsequent medical issues, including chest pain and shortness of breath. This event marked the beginning of her acute medical journey [8].
- **09/12/2022:**
 - The patient was admitted to the emergency department with chest pain and shortness of breath following the accident. Initial assessments included ECG, lab tests, and imaging, revealing rib fractures and signs of pulmonary edema. She was treated with Lasix, gabapentin, Tylenol, and lidocaine patches, and placed on BiPAP [8, 168].
 - A CT Angio Chest was performed, showing cardiomegaly, interstitial pulmonary edema, and rib fractures, but no acute pulmonary embolism [13, 201, 202].
 - The patient was admitted to the ICU for further management, including oxygen therapy [167].
 - Medication orders included apixaban and lisinopril for atrial fibrillation and hypertension [37].
- **09/13/2022:**
 - The patient continued to receive treatment for acute hypoxic respiratory failure, rib fractures, and atrial fibrillation. She remained on BiPAP and was using incentive spirometry [248, 257].
 - A transthoracic echocardiogram showed severe pulmonary hypertension and bi-atrial enlargement [241].
- **09/14/2022:**
 - Progress notes indicated ongoing pain management and efforts to wean the patient off BiPAP [338].
- **09/15/2022:**
 - The patient showed improvement, transitioning from BiPAP to high-flow nasal cannula. Her rib fractures were stable, and her diabetes was managed with insulin [414].

Targeted Summaries

While typical manual summaries maintain neutral focus, automated summaries can target specific issues. Users can prompt the system to act as a senior legal professional answering specific questions: “Summarize events relevant to causation and standard-of-care questions.” This selective reasoning saves enormous time in long, complex files and ensures the output speaks the language of the legal process rather than clinical documentation.

Cross-Witness Analysis: Synthesis and Contradictions

Summarization is only the first step. The greater value lies in synthesis—reading across all transcripts to deliver coherent, defensible answers. Instead of asking “Where does this witness

mention the safety meeting?” counsel can ask “What did all witnesses say about the March safety meeting and identify any contradictions?” The system scans every transcript, compiles responsive testimony, and returns an integrated answer with citations to each source.

LLM-integrated systems like Merlin Alchemy can answer these questions in minutes, running AI-powered searches against hundreds or thousands of transcripts. The system finds the most relevant testimony, synthesizes findings, and provides reports with links to source material for verification.

Key Concerns

Potential Issues of Liability

In reviewing the patient's medical history, several potential issues of liability and malpractice may arise. These concerns primarily revolve around the timeliness and adequacy of care provided, communication barriers, and the management of the patient's complex medical conditions.

Delayed Administration of Medication One potential concern is the delay in administering critical medications. For instance, there was an emergency room delay noted in the administration of Apixaban, an anticoagulant, which was scheduled at 0028 but given at 0205 on 09/12/2022 [57]. Delays in administering anticoagulants can increase the risk of thromboembolic events, particularly in patients with atrial fibrillation, as in this case.

Communication and Language Barriers The patient's primary language is Spanish, and there were multiple instances where language barriers were noted as potential obstacles in care delivery [391, 1258]. Effective communication is crucial in healthcare settings to ensure that patients and their families understand the treatment plans and can make informed decisions. The reliance on family members for translation and the use of interpreters indicate that there might have been lapses in communication, which could lead to misunderstandings or mismanagement of care.

Advance Care Planning and Code Status The patient's code status and advance care planning were significant aspects of her care, with multiple discussions and updates documented [709, 712, 1423]. However, there were instances where the advance directive follow-up was noted as needed, indicating potential gaps in ensuring that the patient's wishes were consistently documented and respected [1469, 2215]. Any failure to adhere to a patient's advance directives could raise legal concerns, especially if the patient receives unwanted interventions.

Management of Complex Medical Conditions The patient's complex medical history, including acute on chronic hypoxemic respiratory failure, severe pulmonary hypertension, and multiple rib fractures, required comprehensive management [1529, 1629]. The records indicate ongoing assessments and adjustments in treatment plans, but the complexity of her conditions necessitates meticulous coordination among healthcare providers. Any lapses in this coordination could lead to suboptimal care, raising concerns about the standard of care provided.

Here is an example from the BP Oil Spill trial showing contradictions in testimony regarding the alleged failure to detect and respond to warning signs:

The ability to ask questions across hundreds of transcripts changes how legal teams approach testimony analysis. Instead of relying on fragmented summaries and memory, attorneys can instantly retrieve every relevant statement on a given topic, compare witnesses, and identify contradictions.

One of the most valuable but tedious transcript tasks is comparing what different witnesses said about the same events. AI can systematically compare testimony, extracting what each witness said, identifying areas of agreement and disagreement, flagging direct contradictions, noting differences in detail or emphasis, and synthesizing comparative analysis.

Here is an anonymized example of one contradiction analysis segment:

Testimony Alleging Failure to Detect and Respond

Expert Richard Heenan (Petroleum Engineer)

Richard Heenan, an expert petroleum engineer, testified that the failure to interpret the negative pressure test correctly was a “gross and extreme departure from the standards of good oilfield practice” ([BP_00000001 §6 ¶57](#)). He stated that the crew missed indications of a well flowing and that the well should have been shut in ([BP_00000002 §16 ¶154](#), [BP_00000002 §16 ¶156](#)).

Expert Calvin Barnhill (Transocean’s Expert)

Calvin Barnhill, Transocean’s own expert, affirmed that the crew missed indications of a well flowing and that the accident could have been averted if a manual flow check followed by an immediate shut-in had occurred after the rig pumps were shut down at 2131 ([BP_00000002 §16 ¶156](#)). He testified that the driller and assistant driller should have monitored the well, observed increasing shut-in drill pipe pressure, and performed a flow check, which would have indicated a positive result and led to shutting in the well ([BP_00000025 §17 ¶819](#), [BP_00000025 §17 ¶824](#)).

Billy Dean Ambrose (Transocean Investigation Lead)

Billy Dean Ambrose, who conducted Transocean’s investigation, acknowledged that an increase in drill pipe pressure at approximately 9:01 PM was a kick indication, though it was not recognized as such at the time ([BP_00000002 §15 ¶150](#)). He stated that between 9:01 PM and 9:08 PM, a significant anomaly occurred with a 100-psi increase and a 40-barrel hydrocarbon influx, which should have been investigated ([BP_00000002 §15 ¶151](#)). By 9:31 PM, an approximately 300-barrel hydrocarbon influx had occurred, indicating a massive kick that had not been detected ([BP_00000002 §15 ¶154](#)).

Ambrose also testified that the failure of the negative pressure test, indicated by anomalous pressure on the drill pipe, should have alerted all monitoring personnel, including the Transocean crew and BP well site leader, to the ineffective cement barrier and communication with the formation ([BP_00000035 §31 ¶318](#), [BP_00000035 §31 ¶320](#)).

The comparison workflow extracts testimony about specified topics from each witness, then identifies statements that confirm each other, contradict each other, add details others omitted, or differ in emphasis or interpretation. The output organizes these comparisons clearly, showing what each witness said alongside each other for easy review. This analysis proves particularly valuable for impeachment preparation and evaluating credibility.

Quality assurance for comparison analysis requires verifying that identified agreements and contradictions are real rather than artifacts of misunderstanding testimony. Sampling proto-

cols check significant contradictions against source transcripts to ensure the AI correctly interpreted what witnesses said.

TOPIC 3: TELEMATICS AND MONITORING

GPS Accuracy Standards

- Bennett testified to 3-meter accuracy (Pinnacle_Summary_0004 §3)
- Miller stated 5-meter accuracy standard (Pinnacle_Summary_0017 §7)

Data Retention Policies

- Bennett: 7-10 days device storage, 30-day system retention (Pinnacle_Summary_0005 §1)
- Martinez: 6-month retention requirement (Pinnacle_Summary_0013 §2)
- Miller: Different retention periods for various data types (Pinnacle_Summary_0017 §4)

Significance Assessment: Inconsistencies in technical specifications and data retention policies could affect the reliability and admissibility of telematics evidence in legal proceedings. These variations may also impact compliance with regulatory requirements for data preservation and accuracy standards.

Practical Impact: Professional Time Focused on Strategy

The transformation from traditional review to AI-assisted analysis delivers measurable operational benefits. Tasks that once required weeks of manual effort complete overnight. Every deposition follows the same analytical template regardless of reviewer, ensuring consistency. Perhaps most significantly, AI summaries can be more comprehensive than their manual counterparts because time constraints no longer force selectivity.

Each conclusion links directly to source citations, providing a clear audit trail—particularly crucial when defending work product to opposing counsel or courts. Organizations preserve insights across matters, avoiding the repeated review that occurs when staff turns over. Most importantly, lawyers and analysts spend their time on argument and strategy rather than transcription and summary.

Like the document review workflows in Chapter 11, transcript and medical record analysis contribute to the continuous intelligence model. Every temporal reference extracted here—witness recollections, clinical timelines, exhibit dates—becomes source material for the comprehensive chronologies examined in Chapter 13. The system maintains cumulative case knowledge rather than treating each discovery phase as isolated and sequential.

Looking Ahead

The capabilities described here—extracting temporal references from testimony, tracking witness statements across depositions, organizing clinical timelines from thousands of pages—all feed into a more comprehensive challenge: constructing integrated chronologies from multiple sources.

Understanding when events occurred, in what sequence, and with what temporal relation-



ships stands at the center of most litigation and investigations. The next chapter examines how AI weaves together testimony, medical records, and documentary evidence into defensible timelines that reveal not just what happened, but when, who knew what at each point, and what contradictions exist across sources. These comprehensive chronologies become the factual foundation for everything that follows—from case strategy through trial presentation.

13

Using GenAI to Build Timelines and Chronologies

CHAPTER THIRTEEN

Using GenAI to Build Timelines and Chronologies

The production arrives on a Tuesday morning—523,000 documents from the defendant’s document management system, email servers, and shared drives. Your team needs to understand the basic chronology of events leading to the regulatory violation alleged in the complaint. When did the company first learn about the contamination? What sequence of decisions led to continuing operations despite warnings? Who knew what, and when did they know it?

A senior associate suggests the traditional approach: identify key custodians, search for obvious date-stamped documents like board minutes and safety reports, review emails chronologically from the likely time period, and manually extract dates into a spreadsheet. She estimates three weeks with two attorneys and a paralegal working full time. That’s 360 hours of professional time before you have even a preliminary timeline.

There’s a better approach.

Building a Timeline from Half a Million Documents

Instead of manual review, you can use AI to systematically extract temporal information across the entire production, identify key events and decision points, organize chronologically while preserving uncertainty, and generate a comprehensive timeline in hours rather than weeks. The workflow uses three core capabilities working in sequence: **Find** to identify relevant documents, **Analyze** to extract temporal information and understand events, and **Answer** to synthesize findings into structured chronologies.

You begin by searching across all 523,000 documents for those discussing the contamination issue. Rather than keyword searches that might miss relevant documents or return thousands of false positives, you use natural language queries: “Documents discussing groundwater contamination or water quality issues at the Springfield facility.” The system identifies 4,200 potentially relevant documents by understanding the concepts regardless of specific terminology used.

These 4,200 documents become your working set for timeline construction. You then use targeted analytical queries to extract temporal information systematically. “When did company personnel first learn about contamination at the Springfield facility?” returns not just documents but synthesized findings: an environmental engineer’s email from March 2022 mentions “elevated readings we discussed last month,” pointing to initial awareness in February. A consultant’s report dated January 15, 2022 describes “recent testing results” suggesting activity in late 2021 or early January.

You continue with a series of focused questions. “When did management receive reports about contamination?” identifies a safety briefing on April 3, 2022 where the VP of Operations was presented with test results. “What decisions were made about facility operations after contam-

ination was discovered?” reveals a May 2022 management meeting where the team decided to continue operations while implementing additional monitoring.

Each query returns relevant documents with the AI extracting temporal references—both explicit dates from document metadata and headers, and implicit dates from phrases like “last month” or “the meeting we had yesterday” that require temporal reasoning to interpret. When an email dated June 15 says “as we discussed in our meeting last week,” the system infers a meeting occurred around June 8.

The system handles varying levels of temporal precision appropriately. A board member testified she learned about the contamination “sometime in mid-2022”—the system represents this as spanning April through July rather than forcing false precision. A consulting engineer’s report has an explicit date of March 22, 2022. An internal memo references testing “scheduled for next week” on February 11, 2022, creating a temporal window of February 14-18.

After extracting temporal information from dozens of analytical queries, you ask for synthesis: “Create a comprehensive timeline of events related to contamination discovery and management response from 2021 through 2023.” The system organizes all extracted temporal information chronologically, identifies where multiple sources reference the same events, flags contradictions that deserve investigation, and generates a structured timeline with full citations to supporting documents.

The entire process—from the initial production to a comprehensive draft timeline—takes eight to twelve hours of attorney time spread across several days. Most of that time is spent reviewing the AI’s findings, verifying critical dates against source documents, and refining queries to fill gaps. The system does the exhausting work of reading hundreds of thousands of documents and extracting temporal references. You do the strategic work of deciding what questions matter and evaluating whether the timeline tells a coherent story.

A Real Example: The BP Deepwater Horizon Timeline

The timeline below shows the critical 90-minute period when rig personnel conducted and misinterpreted the negative pressure test. This chronology was built by analyzing thousands of pages of testimony transcripts, regulatory filings, and internal BP documents. The timeline preserves precise timestamps where evidence supports them while maintaining appropriate uncertainty where witnesses provided approximations.

Negative Pressure Test Interpretation (April 20, 2010, ~16:54-20:00)			
Date/Time	Personnel	Event Summary	Document Reference
April 12-20, 2010	BP Management	BP issued four different negative pressure test procedures with conflicting requirements; final procedure differed from MMS-approved version submitted four days earlier	BP_00001600 Â§2 Â¶19-22; BP_00001418 Â§2 Â¶22-23
April (morning) 20	BP Management	Final temporary abandonment procedure delivered to rig; never received required MMS approval	BP_00001184 Â¶1-2
16:52-16:54	Rig Crew	Critical spacer displacement error: personnel pumped X total strokes instead of 'X strokes of pill, Y strokes of sea water,' positioning spacer across BOP stack instead of 1000 feet above it	BP_00000865 Â¶15; BP_00001552 Â§1 Â¶9; BP_00001552 Â§24 Â¶565
~19:30-19:45	Jason Anderson (Toolpusher), Robert Kaluza (BP Well Site Leader)	CRITICAL DECISION POINT: Drill pipe showed 1,260-1,400 psi while kill line showed ~30 psi; both should have been zero for successful test; Anderson proposed 'bladder effect' explanation, stating 'he had seen that before and that was normal'; Kaluza responded 'that was possible'	BP_00001600 Â§3 Â¶26-27; BP_00001418 Â§3 Â¶27-28; BP_00000049 Â§7 Â¶74-75

Negative Pressure Test Interpretation (April 20, 2010, ~16:54-20:00)			
Date/Time	Personnel	Event Summary	Document Reference
~19:30-19:45	Multiple Personnel (Anderson, Kaluza, Vidrine)	Personnel 'were satisfied that the annular compression or the bladder effect concept explained and answered the question about the differential pressure' and 'Nobody seemed concerned about it'	BP_00000049 Â§7 Â¶77-78
~19:30-19:45	BP Well Site Leaders (Kaluza and Vidrine)	BP well site leaders 'failed to recognize that the well lacked integrity'; 'recklessly accepted' the failed test	BP_00001371 Â§14 Â¶141; BP_00001391 Â§19 Â¶266; BP_00001402 Â§23 Â¶256
~19:45-20:00	Transocean Toolpushers and BP Well Site Leaders	Extensive discussions to resolve conflicting pressure data; Toolpusher proposed 'bladder effect' based on previous observations; all personnel accepted explanation despite it being incorrect	BP_00001397 Â§15 Â¶169, 171
~19:45-20:00	BP and Transocean Personnel	Despite conflicting data showing 1,400 psi on drill pipe and 0 psi on kill line, personnel declared test successful	BP_00001600 Â§3 Â¶27-28, 33; BP_00000001

Several features make this timeline effective. It maintains precise timestamps where evidence supports them. It preserves appropriate uncertainty where witnesses gave approximate times (“around 19:45”). Each entry includes clear attribution to source documents—the testimony transcripts and email identifiers that support each temporal claim. The chronology captures both documented actions (the drill crew observing anomalies) and critical gaps (the period between first concerns and the explosion when the escalating danger wasn’t adequately addressed).

Building this timeline manually would have been prohibitively expensive. AI-assisted timeline construction made it practical to systematically analyze testimony from multiple witnesses, extract every temporal reference, reconcile timing conflicts, and identify the gaps that proved most revealing. Understanding that a critical three-week communication silence occurred during the period when the Board approved a product launch—or that test results showing problems were documented but never discussed in management meetings—often matters more than documenting routine events.

This approach transforms timeline construction from a sampling exercise limited by available attorney time into comprehensive temporal analysis across all available evidence. Let’s examine how these systems work and what discovery professionals need to understand about their capabilities and limitations.

Understanding Temporal Evidence

Discovery materials contain three types of temporal information, each requiring different handling. Explicit dates are straightforward—email timestamps showing “2023-03-15 09:42:13,” meeting minutes dated “March 22, 2023,” contracts with “Effective Date: January 15, 2023.” These dates appear in metadata or document text and need no interpretation.

Implicit temporal references require reasoning. When an engineer writes “we discussed the battery overheating issue last month” in a March 15 email, AI systems calculate this refers to mid-February. A QA report dated February 18 mentioning testing “scheduled for next week” generates a temporal window of February 25-March 3. The CEO testifying she learned about problems “sometime in early 2023” yields a three-month range rather than a forced midpoint.

Temporal contradictions matter most. One witness recalls learning about defects in February. Documents show the first management briefing occurred March 22. This gap creates impeach-

ment opportunities or suggests informal communication that wasn't documented. The timeline should surface these conflicts, not suppress them.

The challenge is preserving appropriate precision. Some events have exact timestamps. Others are approximate—"early 2023" or "around mid-March." Still others are relational—"before we learned about the problem" or "shortly after the Board meeting." Effective timelines handle this variation without forcing false precision that exceeds what evidence supports.

How AI Builds Comprehensive Chronologies

AI systems combine three techniques to construct timelines. Explicit date extraction pulls timestamps from metadata and standard date formats in text. Temporal reference interpretation translates phrases like "last month" or "next week" into date ranges based on the document's own date. An email dated June 15 saying "as we discussed in our meeting last week" generates an inferred meeting date around June 8.

The system produces structured output for each temporal reference: the date or timeframe with precision indicators, the associated event, the source document with citation information, and confidence levels. When sources conflict—Document A dates an event to March while Document B suggests April—the system preserves both rather than forcing artificial consistency.

Integration across source types reveals patterns. Emails show engineers discussing battery concerns starting mid-February, with intensity increasing sharply in mid-March (from 15-20 messages weekly to 60 in one week). Then engineering-to-management communications cease for three weeks, from March 23 through April 12—covering the period when the Board approved launch. Document analysis finds test reports showing continued problems throughout March. Meeting minutes from March 22 document a management briefing. The CEO testifies she learned about concerns "in early 2023." Board minutes from April 3 show launch approval without mentioning safety issues.

This integrated timeline reveals a troubling sequence: problems identified in February, testing confirming issues through March, management briefed in late March, yet Board approval proceeding in early April without documented safety discussion. The timeline doesn't interpret—it makes temporal relationships visible that manual review might miss.

The Value of Gap Analysis

Timeline systems identify not just what's present but what's conspicuously absent. In our product defect case, regular email traffic between engineers and management suddenly stops for three weeks—precisely during the period when the Board approved launch despite documented safety concerns. That silence warrants investigation.

Pattern analysis quantifies what attorneys might sense intuitively. The system might flag: "Email volume about battery testing dropped 85% in the three weeks before Board approval compared to the prior three months, despite test results showing continued problems." Communication routing can shift tellingly—the lead engineer emailed the CEO directly 15-20 times monthly about product development through mid-March, then zero direct emails after March 22 when engineers briefed management about safety concerns.

These patterns raise questions without proving anything. The communication gap might reflect innocent scheduling coincidence, or it might suggest deliberate efforts to avoid documentation before a critical decision. The routing change might represent sensible organizational adjustment, or deliberate insulation of senior executives from bad news. Timeline analysis surfaces these anomalies systematically rather than depending on attorneys to notice patterns while reviewing hundreds of thousands of documents.

Specialized Timeline Applications

Three timeline types address specific litigation challenges. Each reveals different aspects of how events unfolded and who knew what when:

Knowledge timelines prove particularly valuable when they reveal that key decision-makers possessed critical information but failed to act. The Mark Hafle timeline illustrates this precisely. Hafle was BP's shoreside senior drilling engineer. At 8:52 PM on the night of the blowout, he received a call detailing impossible pressure readings from the negative pressure test. His timeline reveals the tragedy of correct knowledge combined with failed intervention.

Mark Hafle, BP Shoreside Senior Drilling Engineer

- ~20:52 (8:52 PM): Hafle received a phone call from Don Vidrine, the BP Well Site Leader on the Deepwater Horizon, to discuss the results of the negative pressure test. Vidrine reported conflicting data: zero pressure on the kill line but pressure on the drillpipe (BP_00001600 §6 ¶59-60, BP_00001418 §7 ¶71).
- ~20:52 (8:52 PM): Hafle correctly identified the technical failure of the test based on the data provided by Vidrine. He stated: "you cant have pressure on the drill pipe and zero pressure on the kill line in test that is properly lined up" (BP_00001600 §6 ¶60, BP_00001418 §7 ¶71). This observation coincided with modeling that indicates the well had begun flowing on its own at this time (BP_00001418 §7 ¶71).
- ~20:52 - 21:02 (8:52 PM - 9:02 PM): Despite his correct technical assessment that the conflicting data "could not indicate successful pressure test," Hafle accepted Vidrine's conclusion that the test had succeeded anyway (BP_00001600 §4 ¶34-35, BP_00001418 §4 ¶36-37).
- ~20:52 - 21:49 (8:52 PM - 9:49 PM): During the approximately one hour between his phone call with Vidrine and the first explosion, Hafle took no action to re-evaluate the test, rerun the test, or implement any well control response (BP_00001600 §6 ¶61, BP_00001418 §1 ¶9, BP_00001418 §7 ¶72).

The timeline was created in seconds from the entirety of the BP liability trial.

Decision Timelines prove particularly powerful when they reveal not just what decisions were made, but the sequence in which information became available and choices were executed. The April 20th decision timeline shows five critical failure points where correct action could have prevented the disaster.

Timeline of Top 5 Key Decisions and Failures: April 20, 2010

1. DECISION TO ACCEPT THE "BLADDER EFFECT" EXPLANATION (~19:30-19:45)

1. What Happened: Faced with impossible pressure readings (1,400 psi on the drill pipe vs. 0 psi on the kill line), rig leaders accepted the scientifically unfounded "bladder effect" theory to explain

the discrepancy ([BP_00000049 §7 ¶¶74-75](#); [BP_00001397 §15 ¶¶169](#)).

2. Why It Was Key: This was the single most critical failure in judgment. It allowed personnel to rationalize away clear evidence of test failure, creating a false consensus that the well was secure and enabling all subsequent mistakes.

2. DECISION TO FORMALLY DECLARE THE NEGATIVE PRESSURE TEST SUCCESSFUL (~20:00)

1. What Happened: A BP Well Site Leader formally declared the second negative pressure test successful, despite the contradictory data ([BP_00001359 §18 ¶¶201](#)).
2. Why It Was Key: This official declaration was the “point of no return.” It provided the final authorization to proceed with the next operational step—displacing the heavy mud with seawater—which directly created the underbalanced condition that allowed the blowout.

3. DECISION TO PROCEED WITH MUD DISPLACEMENT (~20:02)

1. What Happened: Based on the flawed test conclusion, the crew began displacing the heavy drilling mud from the riser with much lighter seawater ([BP_00001552 §16 ¶¶345](#); [BP_00001600 §4 ¶¶40-41](#)).
2. Why It Was Key: This action knowingly created an underbalanced state, removing the primary hydrostatic barrier holding back the reservoir. It transformed a static, failed test into a dynamic and dangerous well condition.

4. DECISION BY SHORESIDE ENGINEER TO NOT INTERVENE (20:52)

1. What Happened: BP shoreside engineer Mark Hafle was told of the impossible pressure readings and correctly stated the test had failed, but then took no action to stop operations ([BP_00001600 §6 ¶¶60-61](#); [BP_00001418 §4 ¶¶36-37](#)).
2. Why It Was Key: This was the last clear opportunity to prevent the hydrocarbon influx entirely. A qualified expert with authority to intervene correctly identified the failure but failed to act at the precise moment the well became underbalanced.

5. DECISION TO IGNORE PRESSURE ANOMALIES DURING DISPLACEMENT (21:01-21:14)

1. What Happened: Personnel failed to investigate a 100-psi pressure increase (21:01-21:08) and, most significantly, a 250-psi increase while pumps were off for a sheen test (21:08-21:14) ([BP_00000002 §15 ¶¶149, 153](#)).
2. Why It Was Key: These were unambiguous signs the well was flowing (a “kick”). The failure to investigate and shut in the well at this stage allowed a manageable influx of ~40 barrels to grow into an uncontrollable one.

Once again, this was created in seconds and could have been expanded to provide a lot more detail.

Communication pattern timelines reveal not just what was said, but who said it to whom, and how information flow changed as events progressed. In the BP case, analyzing communication patterns revealed critical breakdowns in the final hour—particularly the failure to inform the mudlogger when flow was diverted overboard, effectively blinding a key safety monitoring system.

Summary of Key Communications in the Final Hour

1. ~20:52: The Critical Phone Call (Vidrine & Hafle)

1. Who: Don Vidrine (BP Well Site Leader) called Mark Hafle (BP Shoreside Engineer).
2. What: Vidrine reported the impossible pressure readings. Hafle correctly identified that the test had failed, stating “you can’t have pressure on the drill pipe and zero pressure on the kill line in test that is properly lined up.”
3. Outcome: Despite Hafle’s correct technical assessment, both men failed to take action. Vidrine

proceeded with operations, and Hafle did not intervene. This was the last clear opportunity to prevent the influx entirely.

4.Reference: ([BP_00001600 §6 ¶¶59-61](#); [BP_00001418 §7 ¶¶71-72](#))

2. ~21:14: The Decision to Resume Displacement (Vidrine & Driller)

1.Who: Don Vidrine (BP Well Site Leader) communicated with the driller.

2.What: After the sheen test (where a 250-psi pressure increase went unnoticed), Vidrine was in the driller's shack, reviewed the data, and indicated "everything appeared to be fine." He then gave the instruction, "Gave okay to start dumping."

3.Outcome: This decision authorized the crew to continue displacing mud with seawater, despite the well being live and flowing.

4.Reference: ([BP_00000025 §27 ¶¶1333-1335, ¶¶1356-1360](#))

3. ~21:30: The Crew's Discussion of Anomalies (Driller, Toolpusher, & Mr. Young)

1.Who: The driller and toolpusher were discussing the well status in the drill shack. Mr. Young, another witness, overheard them.

2.What: They were discussing "a pressure differential" and the need "to circulate out the well."

3.Outcome: This shows the crew recognized an anomaly. However, their diagnosis was incorrect; they believed it was a circulation or trapped pressure issue, not a major well control event.

4.Reference: ([BP_00000025 §28 ¶¶1413-1415](#))

4. ~21:31: The Flawed Command (Driller & Floorhand)

1.Who: The driller gave an order to a floorhand (Caleb Holloway).

2.What: Instead of shutting in the well, the driller ordered the floorhand to open a valve and "bleed this pressure off."

3.Outcome: This action was a fundamental violation of well control procedure. It treated a kick as a minor pressure issue and allowed the influx to continue escalating, making the situation "too far gone" ([BP_00000025 §29 ¶¶1458](#)).

4.Reference: ([BP_00000025 §29 ¶¶1437-1438](#); [BP_00000305 ¶3](#))

Significant Lack of Communication

The summaries also highlight a critical lack of communication, particularly with the mudlogger, Joe Keith. After 21:08, the decision to divert flow overboard was made without informing him, effectively blinding his sensors to the ongoing influx ([BP_00001396 §9 ¶¶99](#); [BP_00001396 §6 ¶¶64](#)). This breakdown in communication removed a key layer of safety monitoring during the most dangerous phase of the operation.

What Makes These Timelines Possible?

The BP timelines shown above drew from the entire liability trial record—thousands of pages of testimony transcripts from dozens of witnesses, tens of thousands of pages of technical documents, regulatory filings, emails, meeting minutes, and engineering reports. Manual construction of even one of these specialized timelines would have consumed weeks of attorney time. AI-assisted timeline construction made comprehensive analysis practical by systematically processing the entire trial record, extracting every temporal reference regardless of format or precision level, organizing chronologically while preserving uncertainty, and identifying patterns that emerge only from complete coverage. This comprehensive extraction and organization occurred in hours rather than weeks.

But the AI didn't "write" these timelines independently or generate original conclusions. Rather, it extracted temporal information, organized it chronologically, identified patterns and gaps, and presented findings with full citations for attorney verification. Discovery professionals then reviewed the extracted information, verified critical temporal claims against source documents, refined the timeline focus based on litigation strategy, and structured the presentation for specific purposes—trial exhibits, internal work product, or expert witness preparation. The timelines represent human-AI collaboration: the system handles exhaustive extraction and organization, while attorneys provide strategic direction, quality control, and interpretive judgment.

Quality Assurance: The Three Critical Checks

Timeline quality depends on three verification processes. First, accuracy sampling: select 50-100 random entries and verify against source documents that dates are correctly extracted, events accurately described, and precision levels appropriate. Target 95% accuracy. Watch for common errors—date format confusion (day-month versus month-day), timezone miscalculations, and relative date interpretation failures near month boundaries.

Second, completeness review: compare the AI-generated timeline to chronologies legal teams developed through other means. Have attorneys note missing events. Verify all key decision points appear with appropriate dates. Every temporal claim needs a citation—"Management briefed on safety concerns - March 22, 2023" must reference meeting minutes, calendar entries, or testimony establishing that date.

Third, conflict documentation: when sources disagree, preserve both versions rather than forcing consistency. "Engineers briefed on test results - March 22 per meeting minutes; CEO testified this occurred 'sometime in early 2023, possibly as late as April.'" These documented contradictions often prove more valuable than artificially resolved inconsistencies.

Looking Ahead

This chapter examined how AI transforms timeline construction from a manual sampling exercise into comprehensive temporal analysis. Discovery professionals can systematically extract temporal information across entire productions, organize complex chronologies while preserving appropriate uncertainty, identify patterns and gaps that raise questions for investigation, and generate timelines serving diverse purposes from strategic analysis to trial presentation.

But timeline construction represents foundational work, not final product. Understanding what happened when, who knew what at each point, and how events unfolded creates the factual framework supporting all trial advocacy. The next challenge is transforming that chronological understanding into persuasive narratives that convince judges and juries. Chapter 14 examines how Generative AI assists with this transformation—taking the factual foundations built through timeline analysis and marshaling them into compelling opening and closing arguments.

14

Using GenAI to Write Powerful and Persuasive Opening and Closing Arguments

CHAPTER FOURTEEN

Using GenAI to Write Powerful and Persuasive Opening and Closing Arguments

Trial advocacy has always demanded the synthesis of vast amounts of evidence, the ability to construct compelling narratives, and the skill to marshal facts into persuasive arguments. For decades, these tasks required teams of lawyers working countless hours to review transcripts, analyze exhibits, and craft arguments that could sway judges and juries.

Generative AI changes this landscape. Modern Large Language Models can now process thousands of pages of trial materials in minutes, identify the most relevant evidence for specific legal theories, and generate sophisticated first drafts of opening and closing arguments that previously would have taken weeks to develop. This chapter demonstrates how smart discovery professionals can leverage these capabilities to transform trial preparation while maintaining the professional judgment and strategic thinking that defines excellent advocacy.

The question is no longer whether AI can assist with trial preparation. The question is how legal professionals will adapt to these powerful tools and integrate them into their practice to serve their clients more effectively.

The Challenge: Creating Arguments from Complex Trial Materials

Opening and closing arguments represent the pinnacle of trial advocacy. They require lawyers to:

- Synthesize massive volumes of testimony and exhibits
- Identify the most compelling evidence for their theory of the case
- Construct narratives that acknowledge weaknesses while emphasizing strengths
- Anticipate and counter opposing arguments
- Present complex technical information in accessible ways
- Adapt arguments to the specific perspectives of different parties in multi-party litigation

Traditional preparation for these arguments involves teams of lawyers spending days or weeks reviewing trial transcripts, organizing exhibits, drafting argument sections, and revising until the final product emerges. Even with dedicated teams, the time pressure and volume of material make it difficult to ensure that every relevant piece of evidence has been considered and integrated into the argument.

Generative AI transforms this process by enabling comprehensive analysis of the entire trial record in a fraction of the time, while producing sophisticated first drafts that lawyers can refine and personalize.

The BP Oil Spill Trial: A Case Study in AI-Assisted Advocacy

The Deepwater Horizon oil spill trial provides an ideal case study for demonstrating the power of GenAI in trial preparation. On April 20, 2010, an explosion on the Deepwater Horizon drilling

rig killed 11 workers and triggered an oil spill that released approximately 4.9 million barrels of crude oil into the Gulf of Mexico over 87 days. The Phase One trial focused on liability, with multiple defendants facing complex questions of shared responsibility.

The trial generated thousands of pages requiring multiple parties to construct arguments from different strategic perspectives. This complexity makes it an excellent test case for evaluating GenAI's capabilities in legal analysis and argument construction.

Using the Merlin Alchemy platform, we demonstrated how GenAI could analyze this massive trial record and generate sophisticated closing arguments for each party: TransOcean, Halliburton, the government plaintiffs, and BP itself. Each argument required the AI to adapt to different legal positions while maintaining rigorous analytical standards and comprehensive use of the available evidence.

The Methodology: A Systematic Approach to AI-Assisted Argument Development

The process combines AI's analytical power with human legal expertise.

Step One: Document Integration and Repository Creation

The foundation of effective AI-assisted argument development is a comprehensive, well-organized repository of trial materials. This includes:

Trial Transcripts: Complete testimony from all witnesses, organized by date and witness name. The transcripts should be processed to enable both semantic and keyword-based searching.

Trial Exhibits: All documents, communications, technical reports, and other materials admitted into evidence. Each exhibit should be indexed and linked to the testimony that references it.

Document Structuring: Materials should be divided into logical sections that enable the AI to locate and retrieve relevant evidence efficiently. For the BP trial, we divided the materials into over 9,000 sections of testimony and exhibit text.

Metadata Integration: Each document section should include metadata indicating the source, date, witness, and other relevant contextual information that helps the AI understand the significance of the evidence.

This preparation work is essential because the quality of AI-generated arguments depends directly on the quality and organization of the underlying materials.

Step Two: Developing the Strategic Outline

First, create a comprehensive strategic outline. This outline development process leverages the AI's ability to analyze the entire trial record and identify the strongest arguments available. The prompt for outline development should include:

Role Definition: Clear instructions about which party the AI is representing and the overall strategic objective. For example: "You are the lead trial lawyer for British Petroleum in the

Deepwater Horizon Oil Spill Phase One Liability Trial. You have years of experience defending large corporations in complex disasters.”

Strategic Direction: Specific guidance about the theory of the case and key themes to emphasize. For BP’s defense, this included: “Your goal will be to push responsibility for this terrible tragedy to the other defendants and to show that BP should not be held liable if at all possible. Tell the story that BP was not the operator of the well and wasn’t responsible for the key elements that failed here.”

Output Requirements: Clear instructions about the format and content of the outline, including the level of detail needed for each section.

The AI processes the entire trial record through sophisticated algorithmic search methods to identify evidence that supports the requested strategic approach. For the BP trial, we used three complementary search techniques:

- 1. Natural Language Search:** Semantic, vector-based searching that identifies contextually relevant evidence even when specific keywords are not present
- 2. Keyword Algorithm:** Traditional keyword searching to ensure important specific terms and phrases are captured
- 3. Machine Learning Classification:** Trained algorithms that learn from initial seed rankings to identify similar relevant sections

This multi-method approach ensures comprehensive coverage of the trial record, catching evidence that any single method might miss.

The resulting outline provides a detailed roadmap for the argument, organized into major sections with supporting subsections and specific citations to testimony and exhibits. For BP’s closing argument, the outline included seven major sections:

- Introduction: The Tragedy in Context
- BP Was Not the Direct Operator of the Well
- Technical Failures Beyond BP’s Direct Control
- Critical Decision-Making and Operational Failures
- BP’s Safety Culture and Risk Management
- Shared Responsibility Among Multiple Parties
- Conclusion: A Fair Assessment of Responsibility

Each section included detailed subsections with specific document citations, creating a comprehensive framework for the final argument.

Step Three: Generating the Argument Text

For lengthy arguments, this is best done in sections to maintain quality and enable iterative refinement.

The prompt for argument generation should include:

Reinforced Role Definition: Restate the lawyer’s role and expertise to maintain consistency with the outline phase.

Specific Instructions: Clear guidance about format, tone, and objectives. For example: “Make your argument in narrative form as if you were giving it to the Judge in this case. Include citations to testimony and exhibits that support your arguments.”

Outline Section: The specific portion of the outline being addressed in this generation run.
Citation Requirements: Explicit instructions to support all factual assertions with specific citations to the trial record.

For the BP closing argument, we divided the outline into three groupings and generated the argument in three separate runs, each processing the relevant sections against the 9,000+ document sections in the repository. This approach allows the AI to focus on each portion of the argument while maintaining access to the entire trial record for evidence selection.

The AI generates narrative argument text that follows the outline structure, integrates the most relevant evidence from the trial record, and includes specific citations linking each factual assertion to its supporting testimony or exhibit. The result is a sophisticated first draft that reads like an experienced trial lawyer’s argument, complete with rhetorical devices, strategic emphasis, and comprehensive evidence integration.

Step Four: Combining and Refining the Sections

The final technical step is combining the separately-generated sections into a unified argument. Modern LLMs output their text in Markdown format, which can be easily integrated into word processing applications. Google Docs, for example, includes a “paste from Markdown” option that automatically formats the text with proper headers, emphasis, and structure.

Once combined, the argument should be reviewed for:

Narrative Flow: Ensuring smooth transitions between sections and logical progression of arguments.

Consistent Tone: Verifying that the different sections maintain a consistent voice and strategic approach.

Citation Accuracy: Checking that all citations are properly formatted and accurately reference the trial record.

Strategic Emphasis: Confirming that key arguments receive appropriate prominence and that the overall structure effectively supports the theory of the case.

This review process is where human legal expertise becomes essential. The AI provides a comprehensive first draft, but skilled lawyers must refine the argument to reflect their judgment, emphasize their strategic priorities, and ensure that the final product represents their best advocacy.

The Results: Sophisticated Arguments in Hours, Not Weeks

The BP trial series demonstrated remarkable results across multiple parties and strategic approaches, with each argument requiring the AI to adapt its analytical approach to fundamen-

tally different legal positions.

TransOcean's Closing Argument

For TransOcean, the drilling rig operator, the challenge was acknowledging operational responsibility while shifting primary liability to BP as the well owner and lessee. The AI constructed a nuanced argument that balanced these competing demands. The argument acknowledged TransOcean's role as rig operator while emphasizing BP's ultimate authority over critical decisions. It documented specific instances where BP overrode TransOcean's safety recommendations and integrated technical testimony about drilling operations and well control to construct a narrative of shared responsibility with primary fault lying with BP.

Halliburton's Closing Argument

Halliburton's position as cementing contractor required emphasizing their limited contractual scope:

- Clearly delineated contractual boundaries separating Halliburton's work from BP's decision-making authority
- Documented evidence of BP systematically prioritizing cost over safety
- Identified specific examples of BP overriding contractor safety recommendations

Government Plaintiffs' Closing Argument

The government's position required demonstrating gross negligence and willful misconduct while establishing BP's primary liability. The AI developed an argument that synthesized evidence across multiple parties to show coordinated failures, emphasized the environmental and economic devastation caused by the spill, documented repeated warnings ignored by the defendants, and constructed a narrative of reckless decision-making driven by profit over safety. This argument required the most complex evidence synthesis, drawing connections across testimony from all parties to establish the pattern of negligence.

BP's Closing Argument: The Ultimate Challenge

The most difficult assignment was creating BP's defense argument. Most legal observers viewed BP's position as nearly indefensible given the overwhelming evidence of their responsibility. The court ultimately ruled that BP was guilty of gross negligence and willful misconduct, apportioning 67% of blame to BP.

Yet the AI successfully constructed a legitimate defense strategy based on shared responsibility in deepwater drilling operations. The argument:

- Acknowledged the tragedy while establishing the complex nature of deepwater drilling
- Emphasized BP's limited operational control as non-operator of the well
- Highlighted technical failures by other parties, particularly the blowout preventer maintained by TransOcean
- Identified cementing failures by Halliburton
- Documented critical operational decisions made by TransOcean's crew

- Presented evidence of BP's safety commitment and reliance on contractor expertise
- Argued for fair allocation of responsibility among all parties

The final argument spanned 60 pages of sophisticated legal reasoning supported by specific citations to the trial record. The entire argument generation process was completed in approximately two hours, demonstrating the dramatic efficiency gains possible with AI-assisted trial preparation.

Key Benefits of AI-Assisted Trial Argument Development

The BP trial series demonstrated several significant advantages of using GenAI for trial preparation:

Comprehensive Evidence Analysis

The AI processed thousands of pages in minutes, ensuring no relevant evidence was overlooked. The systematic approach catches testimony using different terminology, exhibits referenced in multiple contexts, and connections between witnesses' accounts that human reviewers might miss.

Multiple Strategic Perspectives

The ability to generate arguments for different parties using the same evidence demonstrates how GenAI can help legal teams explore alternative theories of liability and defense strategies. Rather than committing to a single approach early in trial preparation, lawyers can quickly generate and evaluate multiple strategic options.

This flexibility is particularly valuable in complex multi-party litigation where the optimal strategic approach may not be immediately obvious. The AI can generate arguments from different perspectives, allowing the legal team to assess the strengths and weaknesses of each approach before committing resources to full development.

Rapid Iterative Development

Traditional argument preparation follows a linear process: review documents, draft sections, revise, incorporate new evidence, revise again. This process can take weeks, and significant strategic pivots late in the process can require starting over.

GenAI enables rapid iteration. If new evidence emerges or strategic priorities shift, the AI can quickly generate a revised argument incorporating the new direction. This flexibility allows legal teams to remain responsive to developments throughout trial preparation.

Cost Efficiency

Perhaps the most striking advantage is the dramatic reduction in cost. Generating sophisticated first drafts in hours rather than weeks produces substantial savings on legal fees. More importantly, it allows senior lawyers to focus their expensive time on strategic refinement and personalization rather than initial research and drafting.



For the BP arguments, the entire project cost less than a nice dinner for two. While senior trial lawyers would never present AI-generated arguments verbatim, having a comprehensive first draft that would take associates weeks to prepare provides enormous value.

Quality Control and Citation Accuracy

Every point in an AI-generated argument is linked to specific evidence in the trial record. This comprehensive citation enables efficient verification of factual assertions and ensures that the argument is built on the actual trial record rather than assumptions or generalizations.

The systematic approach also helps ensure consistency in how evidence is characterized and reduces the risk of inadvertent misrepresentation that can occur when multiple team members work on different sections without complete coordination.

Practical Applications Beyond Major Trials

The same methodology applies to routine civil litigation, criminal defense, appeals, and arbitrations—analyzing depositions, police reports, trial records, and documentary evidence to generate arguments in a fraction of traditional time.

Limitations and Considerations

AI can analyze evidence and generate arguments, but it cannot exercise professional judgment about strategic priorities, assess the credibility implications of different argument choices, or evaluate how arguments will resonate with specific judges or juries. These remain fundamentally human skills.

The AI-generated argument is a first draft, not a final product. Experienced trial lawyers must review, refine, and personalize the argument to reflect their strategic judgment and advocacy style.

Ethical Obligations Apply

Lawyers remain responsible for the accuracy and propriety of all arguments they present, regardless of how those arguments were prepared. This includes:

- Verifying that all citations accurately represent the evidence
- Ensuring that characterizations of testimony are fair and not misleading
- Confirming that legal arguments are supported by applicable law
- Maintaining client confidentiality when using AI platforms
- Complying with all court rules regarding disclosure of AI assistance

Context and Nuance Matter

AI can miss subtle contextual factors that experienced lawyers immediately recognize. This includes understanding the relative credibility of different witnesses, appreciating the importance of demeanor evidence that doesn't appear in transcripts, and recognizing tactical implications of emphasizing certain arguments over others.

Human review must ensure that the AI-generated argument reflects appropriate contextual judgment.

Repository Quality Determines Output Quality

The AI can only work with the materials it is given. Incomplete document repositories, poor document organization, or inadequate metadata will result in inferior arguments. Significant upfront investment in proper document preparation is essential for optimal results.

Looking Forward: The Future of Trial Advocacy

The question for legal professionals is not whether AI will transform their practice, but how quickly they will adapt to leverage these powerful tools. Forward-thinking attorneys who embrace AI capabilities while maintaining professional judgment and strategic thinking will deliver better results for their clients while working more efficiently.

GenAI amplifies legal expertise rather than replacing it. Trial advocacy will always require the human qualities that define great lawyers: strategic thinking, persuasive communication, ethical judgment, and dedication to clients. GenAI simply provides better tools for applying those qualities to serve clients more effectively. Smart discovery professionals who master these tools will be better positioned to deliver outstanding results in an increasingly competitive legal marketplace.

Conclusion

This chapter demonstrated how Generative AI transforms trial preparation—analyzing complex litigation materials and producing compelling opening and closing arguments that integrate comprehensively with underlying evidence. The BP Oil Spill trial series showed that AI can generate sophisticated arguments from multiple strategic perspectives.

Chapter 15 examines how individual AI capabilities integrate into comprehensive AI-augmented discovery practice. We'll explore how organizations build institutional capabilities supporting AI adoption, what skills and competencies discovery professionals need as AI capabilities advance, how to measure success and ROI across AI-augmented workflows, and where this technology is heading. The chapter synthesizes lessons from the applications we've examined, moving beyond viewing AI as isolated tools toward understanding it as fundamental enhancement to how discovery professionals work—augmenting capabilities, accelerating workflows, and enabling analyses that would be impractical through traditional methods.

15

Building AI-Augmented Discovery Practice

CHAPTER FIFTEEN

Building AI-Augmented Discovery Practice

Trial advocacy has always demanded the synthesis of vast amounts of evidence, the ability to construct compelling narratives, and the skill to marshal facts into persuasive arguments. For decades, these tasks required teams of lawyers working countless hours to review transcripts, analyze exhibits, and craft arguments that could sway judges and juries.

The goal is moving beyond viewing AI as a collection of isolated tools toward understanding it as fundamental enhancement to how discovery professionals work—augmenting capabilities, accelerating workflows, and enabling analyses that would be impractical through traditional methods.

Essential Competencies for AI-Augmented Discovery

As AI capabilities advance, discovery professionals need evolving competency sets balancing technical understanding with enhanced professional judgment. These competencies fall into three categories: technical literacy enabling effective deployment, professional judgment that AI cannot replicate, and adaptive capacity for continuous learning.

Technical Literacy

Discovery professionals need practical understanding of AI capabilities and limitations without requiring deep technical expertise. This means grasping how large language models work conceptually—enough to make informed decisions about deployment, recognize hallucination risks, understand context window constraints, and appreciate both strengths and current limitations.

This literacy extends to vendor evaluation. Discovery professionals should know what questions to ask about AI features, how to assess whether claimed capabilities actually deliver value, what assurances about accuracy and reliability are meaningful, and what infrastructure AI deployment requires. When vendors claim “AI-powered” features, professionals with adequate literacy can distinguish genuine capabilities from marketing.

Technical understanding develops through education and hands-on experience with AI systems.

Professional Judgment and Synthesis

As AI handles more routine analysis, human value increasingly concentrates in synthesis, judgment, and strategic thinking. Discovery professionals must integrate findings across multiple AI analyses into coherent understanding, recognize patterns and relationships AI might identify separately but not connect, evaluate conflicting information from different sources, and construct narratives explaining how events unfolded based on disparate evidence.

Strategic thinking becomes more critical as AI's analytical capabilities grow. Discovery professionals determine what investigations are worth pursuing given available resources, decide when gathered information suffices for strategic decisions versus when additional investigation is needed, evaluate what theories evidence supports versus what theories should be abandoned, and advise clients on risks and opportunities. The discovery professional's comparative advantage shifts from processing information to interpreting it wisely.

Verification skills represent another essential professional competency. With AI handling substantial analysis, discovery professionals need systematic approaches to verifying outputs without rechecking everything manually. This means developing sampling strategies appropriate to stakes and AI reliability, recognizing red flags suggesting outputs deserve closer scrutiny, tracing AI conclusions to source evidence efficiently, and conducting targeted verification on high-risk areas rather than comprehensive rechecking.

Early AI adoption typically involves intensive verification that becomes more targeted as reliability is demonstrated. A discovery professional might initially verify 100% of AI privilege assessments, then reduce to 25% sampling after accuracy is established, but return to intensive verification when new document types emerge.

Communication and Continuous Learning

Discovery professionals must explain AI-assisted work clearly to diverse audiences—clients who need to understand how conclusions were reached, courts that may question methodology, opposing parties who might challenge AI use, and colleagues who need to build on AI-assisted work. This requires explaining AI capabilities and limitations in accessible terms, describing verification approaches, acknowledging uncertainty appropriately, and relating AI methodology to familiar concepts.

Adaptive capacity—the ability to learn continuously, adjust to changing tools and practices, and maintain effectiveness despite disruption—may be most important for long-term success. Discovery professionals need intellectual humility recognizing that today's best practices might not remain optimal, openness to experimentation despite uncertainty about outcomes, resilience maintaining effectiveness despite setbacks, and growth mindset viewing challenges as learning opportunities.

Integrated Discovery Workflows

Traditional discovery workflows move linearly through stages—collection, processing, review, analysis, and production. AI-augmented workflows remain sequential in some respects but enable iteration and integration across stages that weren't previously practical.

From Sequential Stages to Unified Workflows

Traditional discovery treats each analytical task as a separate sequential stage—first collection, then processing, then privilege review, then issue coding, then relevance assessment, with timeline construction and relationship mapping as follow-on projects after review completes. AI-augmented workflows break down these artificial barriers.

Consider how integrated AI transforms this linear process. Initial assessment uses AI to rapidly

process document samples, understanding population characteristics including key custodians, relevant time periods, major topics, and document types. This assessment informs search and culling strategy while identifying custodians deserving priority review.

Document analysis then proceeds with AI performing multiple analyses in a single pass. Rather than routing documents through separate privilege review, issue tagging, categorization, and metadata extraction stages, a single document receives all these analyses simultaneously.

Timeline construction draws on document analysis results, extracting temporal information systematically as review progresses. As new documents are processed, timelines update automatically. This dynamic approach means understanding of event chronology develops progressively rather than waiting until document review completes. Production review applies these same analytical techniques to opposing parties' materials, using capabilities that assisted with analyzing your own documents.

Iterative Investigation

Beyond linear integration, AI enables iterative investigation where findings from one analysis inform subsequent investigations, creating cycles of discovery and understanding. An investigation might begin with broad questions about key events. Initial AI-powered searches retrieve relevant materials. Analysis reveals key participants, critical timeframes, and unexpected issues. These findings inform targeted follow-up—focusing on identified participants during relevant timeframes, searching for newly identified issues, examining communications earlier analysis suggested should exist.

The result is discovery that adapts dynamically to evolving understanding rather than following rigid plans based on incomplete initial information.

Multi-Source Synthesis

Comprehensive case understanding requires synthesizing information across documents, testimony, timelines, relationships, and other evidence types. AI excels at this synthesis, combining insights from diverse sources into coherent narratives. A synthesis workflow might combine document-based timelines showing when things happened according to contemporaneous records with testimony-based timelines showing when witnesses claim things happened, relationship analysis showing who was involved and how they communicated, gap analysis identifying what's missing from the evidentiary record, and comparative analysis highlighting where sources agree or conflict.

This synthesis reveals patterns invisible when sources are examined separately. Documents might show executives discussing problems in March while testimony claims they didn't learn about issues until May. Relationship analysis might reveal extensive communications about problems during the disputed April period. Gap analysis might identify missing documents from key players during critical times. Synthesizing these findings paints a comprehensive picture of what actually happened versus what parties claim happened.

The synthesis also identifies what additional investigation would be most valuable. When all sources converge on certain facts, those facts seem established. When synthesis shows conflicts across sources, those conflicts deserve investigation. When synthesis identifies gaps

where key information should exist but doesn't, those gaps justify targeted discovery efforts.

Building Organizational Capabilities

Infrastructure and Platform Selection

Organizations need infrastructure supporting AI-augmented workflows. This infrastructure combines existing discovery platforms with AI capabilities, either through platforms with built-in AI features, third-party AI tools integrating via APIs, or custom-developed solutions for specialized needs.

Platform selection involves fundamental tradeoffs. Integration with existing discovery infrastructure minimizes disruption and training requirements. Scalability determines whether platforms can handle typical matter volumes without performance degradation. Security and confidentiality protections must meet professional obligations, particularly for single-tenant architectures protecting client data. Cost structures should align with usage patterns—matter-based pricing works well for episodic use while subscription models suit high-volume practices.

Most critically, platforms should provide transparency into how AI analysis works and what decisions AI makes. Black box systems resist the verification that ethical AI use requires. Discovery professionals should trace AI conclusions to source evidence.. Technical infrastructure also includes data management capabilities for storing and organizing the large volumes of documents, analysis results, timelines, relationship data, and other information AI-augmented workflows generate.

Workflow Design and Integration

Effective AI integration starts with mapping current workflows to understand existing practices, identify pain points, and recognize where AI could add value. Not every workflow step benefits from AI augmentation. Effective integration deploys AI where it provides clear advantages while preserving human control over matters requiring professional judgment.

The critical design question involves defining handoffs between AI and human work. When does AI analysis pass to human verification? What triggers human intervention during AI investigations? Where do humans provide input or direction to AI processes? Clear handoff definitions prevent both gaps where work falls through cracks and redundancy where AI and humans duplicate efforts.

Process design should accommodate iteration and learning. Initial AI deployments might require intensive human oversight. As organizations gain experience and confidence, oversight can lighten for routine applications while remaining intensive for novel uses or high-stakes matters. A firm might initially require attorney review of all AI-generated issue tags, then move to sampling-based verification after accuracy is demonstrated, but return to intensive review when deploying AI for new issue types.

Documentation of AI-augmented processes provides audit trails supporting defensibility if methods are challenged, enables training of new staff, facilitates quality improvement, and demonstrates professional competence if ethical obligations are questioned.

Training and Continuous Improvement

Staff at all levels need training on working effectively with AI-augmented workflows. Attorneys need to understand what AI can reliably do and what remains uncertain, how to verify AI outputs appropriately, and how to explain AI-assisted work to clients, courts, and opposing parties. Case managers need deeper technical knowledge about configuring AI workflows, monitoring performance, and troubleshooting unexpected results. Document reviewers need practical skills for understanding what AI has already analyzed and recognizing when AI assessments seem incorrect.

Continuous improvement uses performance data to refine AI deployments. Organizations might adjust prompts and configurations based on error patterns, modify verification intensity based on demonstrated reliability, expand applications that prove valuable, and retire approaches that don't justify costs. AI deployment should be treated as an ongoing learning process rather than a one-time implementation. Quality management and improvement require leadership commitment—dedicating resources to monitoring, providing mechanisms for feedback, taking corrective action when problems emerge, and rewarding learning from both successes and failures.

The Trajectory of AI in Discovery

Understanding where AI in discovery is heading helps discovery professionals position themselves advantageously as capabilities continue advancing.

Near-Term Developments

Several developments seem probable within the next 2-3 years. Agentic capabilities will mature beyond current implementations, with more sophisticated planning and reasoning enabling complex investigations with minimal human direction, better error correction allowing systems to recognize and fix mistakes autonomously, and improved self-verification where AI systems check their own work before presenting to humans.

Real-time analysis may become practical where AI processes documents as they're collected rather than batch processing. This enables truly continuous case assessment, dynamic timelines and relationship maps updating automatically as evidence emerges, and proactive alerting when important patterns or documents are identified. Instead of waiting until document collection completes to begin analysis, teams would have evolving understanding throughout the discovery process.

Natural interaction will improve as AI systems understand more nuanced instructions and questions. Conversational interfaces will increasingly replace structured queries, systems will clarify ambiguous requests rather than guessing intent, and adaptive interfaces will learn user preferences and adjust accordingly. Discovery professionals might ask "Show me communications about the product failure between engineering and management in the three months before recall" and receive precisely relevant results without constructing complex search queries.

Fundamental Uncertainties

Despite clear trajectories, fundamental uncertainties remain about AI's ultimate role in discovery. Will AI capabilities plateau or continue advancing toward general intelligence that would transform legal practice beyond recognition? Current consensus suggests continued advancement but probably not rapid achievement of artificial general intelligence within the next decade.

How will courts and regulators respond to increasingly sophisticated AI in discovery? Will they embrace capabilities enabling more thorough discovery at lower cost, or impose restrictions limiting AI use? Current signs suggest general acceptance with appropriate oversight, but regulatory environments could evolve. Discovery professionals should monitor developments in jurisdictions where they practice.

The Augmented Discovery Professional

The future of discovery isn't AI replacing discovery professionals—it's discovery professionals augmented by AI capabilities working together more effectively than either could work alone.

Effective AI-augmented practice requires discovery professionals who understand enough about AI to deploy it appropriately while maintaining the professional judgment, ethical grounding, and strategic thinking that AI cannot replicate. It requires continued learning as AI capabilities evolve and as experience reveals what works in practice.

The chapters in Part Three demonstrated how AI transforms specific discovery workflows. This concluding chapter examined how those transformations integrate into comprehensive AI-augmented practice. Organizations and professionals embracing these capabilities will find themselves increasingly effective at serving clients, advancing justice, and building successful practices in an AI-augmented future.

We wish you good luck in what is sure to be an exciting career.



APPENDIX ONE

GEN-AI TERMS SMART LEGAL PROFESSIONALS SHOULD KNOW

APPENDIX ONE

GenAI Terms Smart Legal Professionals Should Know

Understanding the terminology of Generative AI is essential for legal professionals navigating this transformative technology. This glossary provides clear, practical definitions of key concepts, focusing on terms most relevant to discovery, investigation, and legal analysis. The terms are organized alphabetically within categories for easier reference.

Core Generative AI Concepts

Agent / AI Agent An autonomous AI system that can plan, execute, and iterate on tasks without continuous human intervention. Unlike simple chatbots that respond to single prompts, agents can break down complex assignments into sub-tasks, use multiple tools, evaluate their own progress, and adjust their approach based on results. In legal discovery, an agent might independently search documents, identify gaps in the record, formulate follow-up queries, and synthesize findings into a coherent narrative.

API (Application Programming Interface) A set of protocols, routines, and tools for building software applications. In the context of Generative AI, APIs allow users to interact with and access the capabilities of large language models through a defined set of commands and inputs. Discovery platforms use APIs to integrate LLM capabilities into their workflow tools, enabling seamless access to AI analysis without leaving the primary review interface.

Attention Mechanism A technique in neural networks that allows the model to focus on specific parts of the input when generating each part of the output. This is what enables LLMs to understand context and relationships between words separated by long distances in a document. In legal applications, attention mechanisms help the system identify connections between facts mentioned in different sections of lengthy contracts or scattered across multiple depositions.

Chain-of-Thought (CoT) A prompting technique where you instruct the AI to show its reasoning process step by step before providing a final answer. This approach dramatically improves accuracy for complex reasoning tasks. In legal contexts, you might prompt: “Analyze whether this contract creates a fiduciary duty. Show your reasoning step by step, identifying the relevant contractual language, applying the legal standard, and then reaching your conclusion.”

ChatGPT A web-based application offered by OpenAI that allows users to interact with GPT models through a conversational interface. ChatGPT is not the same as GPT itself—it’s a front-end gateway that makes the underlying GPT models accessible through natural language conversation. Think of it as the user interface rather than the engine itself.

Context Window The amount of text an LLM can process and consider at one time, measured in tokens. Modern models have dramatically expanded context windows, with some exceeding 200,000 tokens (roughly 150,000 words or 300 pages of text). This capacity affects the model’s ability to maintain coherence and relevance across long documents or conversations. In

discovery, larger context windows mean you can analyze longer depositions, contracts, or sets of related emails without breaking them into smaller chunks.

Embedding / Vector Embedding A numerical representation of text, images, or other data in a high-dimensional space where semantically similar items are located close together. Embeddings allow AI systems to understand meaning and find conceptually related content even when exact words don't match. In legal search, embeddings enable you to find documents about "terminating the agreement" even if the actual language used was "ending the contract" or "canceling the deal."

Few-Shot Learning Providing the model with a few examples of the desired output format or reasoning pattern before asking it to perform a similar task. For instance, you might show the model two or three examples of how you want deposition testimony summarized, then ask it to summarize additional depositions in the same format. This technique is particularly effective when you need consistent output across multiple similar tasks.

Fine-Tuning The process of taking a pre-trained model and training it further on a specific dataset to adapt it to a particular domain or task. A law firm might fine-tune a model on its own documents and past work product to make it more effective at analyzing cases in their practice area. Fine-tuning is more resource-intensive than prompting but can produce superior results for specialized, high-volume tasks.

Foundation Model A large-scale AI model trained on vast amounts of diverse data that can be adapted to perform many different tasks. GPT, Claude, and Gemini are all foundation models. Rather than training separate AI systems for document review, contract analysis, and deposition summarization, legal professionals can use a single foundation model for all these tasks through appropriate prompting and configuration.

Generative AI (GenAI) A type of artificial intelligence that can generate new content—whether text, images, music, code, or other 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, and structures in the data. Generative AI doesn't simply retrieve existing content; it synthesizes new content that didn't exist before.

GPT (Generative Pretrained Transformer) A form of Generative AI designed to understand, process, and generate human-like text based on the input it receives. The term was coined by OpenAI but has become somewhat generic, used to describe the architecture underlying most modern large language models. As a legal professional, think of GPT as an advanced analytical assistant capable of helping with complex reading, analyzing, and writing tasks.

Grounding Connecting an LLM's responses to specific source material or factual databases to reduce hallucinations and improve accuracy. When an AI system is properly grounded, it bases its answers on actual documents in your collection rather than its general training data. In legal applications, grounding ensures that the system's analysis of your case is based on your specific documents, not on general legal principles or, worse, fabricated information.

Hallucination An AI phenomenon where the model generates plausible but factually incorrect information. This occurs because LLMs are trained to produce coherent, contextually appropriate text, not to verify truth. A model might confidently cite a case that doesn't exist or attribute statements to a witness who never made them. This is particularly concerning for legal

work, which is why verification of AI-generated content against source documents is essential.

Instruction Tuning Training an AI model specifically to follow instructions and respond helpfully to user requests. Models like ChatGPT and Claude have undergone extensive instruction tuning to make them better at understanding and executing user commands. This is why modern LLMs can follow complex, multi-part instructions like “Analyze these contracts, identify differences from our standard form, present findings in a table, and highlight potential risks in bold.”

Large Language Model (LLM) GenAI systems like GPT, Claude, Gemini, Llama, and hundreds of others that are specifically designed to understand, generate, and interact with human language. These models are “large” both in terms of their neural network architecture (billions of parameters) and the volume of data on which they’ve been trained (trillions of words). For legal professionals, LLMs represent powerful tools for analyzing documents, drafting content, and extracting insights from large text collections.

Model In AI contexts, a model is the mathematical representation of patterns learned from training data. When we say “GPT-4” or “Claude Sonnet 4,” we’re referring to specific models—distinct versions with different capabilities, training, and characteristics. Different models excel at different tasks, which is why some platforms offer multiple model options for different use cases.

Multimodal AI AI systems that can process and generate multiple types of content—text, images, audio, video, or code—rather than just one type. Multimodal models can analyze scanned documents, photographs of evidence, charts in financial reports, or audio recordings of calls. This capability is increasingly important for legal work that involves diverse evidence types beyond pure text.

Neural Network / Neural Network Architecture The structure and organization of an artificial neural network, consisting of interconnected nodes (neurons) arranged in layers. This architecture allows the network to learn and process information by adjusting the strength of connections between nodes based on input data and desired output. Neural networks are inspired by biological brains but work quite differently. Understanding that they’re pattern-matching systems rather than reasoning systems helps set appropriate expectations for their capabilities.

Parameter Individual learned values within a neural network that determine how it processes information. Modern LLMs contain billions or even trillions of parameters. More parameters generally mean greater capability and nuance, but also increased computational requirements. When you hear about “GPT-4” versus smaller models, parameter count is one key difference affecting their sophistication and accuracy.

Prompt / Prompting The initial input or instruction given to a Generative AI model to elicit a specific response or output. Prompts can range from simple questions to complex, multi-part instructions with detailed context and formatting requirements. In legal applications, effective prompting is the primary skill that determines how useful AI tools become. A well-crafted prompt provides context, specifies exactly what information you need, and indicates how you want results formatted.

Prompt Engineering The practice of carefully designing prompts to achieve desired outputs

from AI models. This involves understanding how models interpret instructions, what level of detail produces optimal results, and how to structure requests for complex tasks. For legal professionals, prompt engineering is a practical skill—learning how to frame questions and structure instructions to get the most accurate, useful, and relevant responses from AI systems.

Temperature A parameter that controls the randomness or creativity of an AI model’s outputs. Low temperature (near 0) produces more deterministic, focused responses. High temperature produces more creative, varied responses but with greater risk of irrelevant content. For factual legal analysis, you typically want low temperature. For creative tasks like brainstorming case theories, higher temperature might be appropriate.

Token The basic unit of text that language models process. A token might be a word, part of a word, or a punctuation mark. “Discovery” is one token, “investigation” is one token, but “investigational” might be two tokens. Context windows and pricing are typically measured in tokens rather than words. Understanding that roughly 750 words equals about 1,000 tokens helps you estimate document sizes relative to context window limits.

Training Data The vast corpus of text, images, and other content used to train an AI model. Training data shapes what the model knows, how it writes, and what biases it might reflect. Legal professionals should understand that general-purpose models are trained on broad internet content, not primarily on legal materials. This is why domain-specific tools or fine-tuning can improve performance on legal tasks.

Transformer The neural network architecture underlying most modern language models. The “T” in GPT stands for Transformer. Transformers use attention mechanisms to process input sequences, which makes them particularly effective at understanding context and relationships in language. While the technical details are complex, knowing that Transformers are the foundation of modern LLMs helps you understand why they’re so much more capable than earlier AI approaches.

Zero-Shot Learning Asking a model to perform a task without providing any examples, relying solely on the instructions in your prompt. For instance: “Summarize this deposition.” The model hasn’t been explicitly trained on your deposition format but can generalize from its training to understand what you want. Most everyday use of LLMs involves zero-shot learning—you ask questions and give instructions without formal training or examples.

Discovery and Legal-Specific Terms

Conceptual Search Search based on meaning and concepts rather than exact keyword matches. Using embeddings and semantic understanding, conceptual search finds documents related to your query even when they use different terminology. Searching for “termination clauses” might surface documents discussing “exit provisions,” “separation agreements,” or “contract dissolution.” This is fundamentally different from Boolean search and dramatically improves recall of relevant documents.

Document Intelligence AI systems that can extract, understand, and analyze information from documents beyond simple text extraction. Document intelligence can parse complex layouts, understand tables and charts, recognize relationships between sections, and identify document types. In discovery, this means AI can analyze scanned contracts, financial state-

ments with embedded charts, or technical specifications with diagrams—not just clean text files.

Entity Extraction / Named Entity Recognition (NER) Automatically identifying and extracting specific types of information from text, such as names of people, organizations, dates, monetary amounts, or locations. In discovery, entity extraction can quickly compile lists of all individuals mentioned across thousands of documents, identify all references to specific companies or transactions, or extract dates relevant to a timeline. Advanced systems can also identify custom entity types like “contractual obligations” or “compliance violations.”

Issue Coding Categorizing documents according to relevant legal issues, claims, defenses, or topics. AI-assisted issue coding can automatically tag documents with multiple relevant issues, dramatically accelerating privilege review, relevance determination, and work product organization. Unlike simple keyword tagging, AI-based issue coding understands context and can recognize that a document discusses “breach of contract” even when it never uses that exact phrase.

Predictive Coding / Technology-Assisted Review (TAR) Machine learning approaches to document review where the system learns from attorney decisions about relevant documents and then predicts relevance for unreviewed documents. Traditional TAR uses statistical models; newer approaches integrate LLMs for improved accuracy and transparency. TAR can dramatically reduce the number of documents requiring manual review while maintaining or improving identification of relevant material.

Privilege Log Generation Automatically creating privilege logs that describe withheld documents. AI systems can analyze potentially privileged documents, determine the privilege type, identify relevant parties and subject matter, and generate draft log entries following court-required formats. This automation can save hundreds of attorney hours on large matters while reducing the risk of inconsistent or inadequate descriptions.

Relationship Mapping Analyzing documents to identify and visualize relationships between people, organizations, transactions, or events. AI can trace how individuals are connected through email communications, identify patterns in corporate relationships across contracts, or map the flow of decision-making through meeting notes and memos. This capability is particularly valuable in complex investigations, white-collar cases, or multi-party litigation.

Retrieval-Augmented Generation (RAG) A technique where an AI system first retrieves relevant documents from a database, then uses those specific documents to generate its response. This grounds the AI’s output in actual source material rather than its general knowledge. Most legal AI applications use RAG to ensure that document analysis is based on your actual case materials. When a discovery system cites specific documents in its response, it’s using RAG.

Semantic Clustering Automatically grouping documents by conceptual similarity rather than keywords. The system identifies documents that discuss similar topics or themes even when they use different terminology. In discovery, semantic clustering can organize thousands of documents into coherent groups like “contract negotiations,” “performance issues,” or “communications with outside counsel,” making it easier to understand what’s in the collection and find related materials.

Summarization / Document Summarization AI-generated concise summaries of longer doc-

uments or sets of documents. Extractive summarization pulls out key sentences from the original. Abstractive summarization generates new text that captures main points in more concise language. For legal work, you might request extractive summaries to ensure fidelity to original language or abstractive summaries when you need more readable overviews. Most modern systems can do both.

Advanced AI Techniques and Architecture

Agent Orchestration / Multi-Agent Systems Coordinating multiple specialized AI agents to work together on complex tasks. One agent might handle document retrieval, another performs analysis, a third synthesizes findings, and a coordinator manages the workflow. This architecture mirrors how legal teams divide work among associates with different specialties. Multi-agent systems can handle more complex investigations than single agents while maintaining transparency about what each component is doing.

Agentic AI / Agentic Systems AI systems capable of autonomous action to achieve objectives, including planning strategies, executing multiple steps, using various tools, evaluating results, and iterating based on what they learn. Unlike traditional AI that responds to single prompts, agentic systems can conduct multi-step investigations independently. An agentic discovery system might search documents, identify gaps, formulate new queries, analyze results, and produce a comprehensive report—all from a single initial instruction.

Autonomous Reasoning AI systems that can independently break down complex problems, consider multiple approaches, evaluate options, and reach conclusions through multi-step logical processes. In legal contexts, this might involve analyzing whether conduct violates a statute by independently identifying relevant statutory elements, mapping facts to each element, considering exceptions, and reaching a reasoned conclusion without step-by-step human guidance.

Cognitive Architecture The structured organization of AI systems that mimics aspects of human cognition, including memory systems, attention mechanisms, reasoning processes, and decision-making frameworks. Advanced legal AI systems might have separate components for legal knowledge, case-specific facts, reasoning procedures, and output generation, organized in ways that parallel how human lawyers think through problems.

Continual Learning / Lifelong Learning AI systems that can update their knowledge and improve their performance based on new information and feedback without complete retraining. In legal contexts, a continual learning system might improve its understanding of your firm's document coding standards based on attorney corrections, or learn new legal developments from recently decided cases. Note that most current LLMs do not have true continual learning—they're static until explicitly updated by their creators.

Reasoning Trace / Audit Trail A detailed record of an AI system's decision-making process, showing what information it considered, what reasoning steps it followed, and how it reached its conclusions. In legal applications, reasoning traces are crucial for understanding why an AI system classified documents particular ways, what facts it relied on for analysis, or how it reached legal conclusions. This transparency is essential for professional responsibility and quality control.

Reflection / Self-Critique The ability of AI systems to evaluate their own outputs, identify po-

tential errors, and refine their responses. Advanced systems might generate an initial answer, then critique it for accuracy and completeness before presenting a final response. In legal discovery, reflection might involve the system generating a document summary, then checking whether it inadvertently omitted important facts or mischaracterized key points.

Retrieval System / Semantic Search The component of AI systems that finds relevant documents or information from large collections based on meaning rather than keywords. Modern retrieval systems use embeddings to understand queries and documents semantically, enabling them to find relevant material even when exact terms don't match. The quality of retrieval significantly impacts the quality of AI analysis—the system can only analyze what it successfully retrieves.

Tool Use / Function Calling The ability of AI systems to access and use external tools, databases, or APIs beyond their core language capabilities. An AI agent might use a legal research database to find relevant case law, query a document management system to retrieve specific files, or call a calculation tool to analyze financial data. Tool use dramatically expands AI capabilities beyond text generation to include interaction with existing systems and data sources.

Vector Database A specialized database optimized for storing and searching embeddings (vector representations of text or other content). Vector databases enable semantic search across large document collections by finding items whose vector representations are close to a query's vector. Most legal AI platforms with conceptual search capabilities use vector databases to enable fast, accurate semantic retrieval across millions of documents.

Practical Implementation Terms

Batch Processing Analyzing multiple documents or performing multiple tasks in a single automated run rather than one at a time. In discovery, you might batch process thousands of documents for privilege determination, issue coding, or summarization. Batch processing is more efficient than iterative processing but requires careful prompt engineering to ensure consistent quality across all items.

Confidence Score / Confidence Level A numerical measure of how certain an AI system is about its prediction or classification. When reviewing AI-suggested issue codes or relevance determinations, confidence scores help you prioritize which decisions to verify. High-confidence predictions are more likely accurate; low-confidence items deserve closer human review. Understanding confidence scores helps you allocate review resources efficiently.

Corpus / Document Corpus The complete collection of documents being analyzed. In discovery, your corpus might be all documents produced in a matter, all contracts of a certain type, or all depositions. When setting up AI analysis, clearly defining your corpus ensures the system searches the right materials and doesn't miss relevant documents or waste time on irrelevant ones.

Deployment / Model Deployment Implementing an AI system in a production environment where it's actively used for real work rather than just testing. Deployment involves configuring the system, training users, establishing quality control procedures, and integrating with existing workflows. For law firms, deploying AI capabilities requires technical implementation plus change management to ensure attorneys use the tools effectively.

Human-in-the-Loop / HITL Systems where AI performs initial analysis but humans review and can correct AI decisions, with those corrections improving future AI performance. Most responsible legal AI implementations are human-in-the-loop—the AI accelerates work but doesn't make final determinations without attorney oversight. This approach balances efficiency gains with professional responsibility and quality control.

Inference / Model Inference The process of using a trained AI model to analyze new data and generate predictions or outputs. When you submit a document for AI analysis, the system performs inference—applying its learned patterns to this new document. Inference is distinct from training; it's what happens when you actually use an AI system for productive work.

Latency / Response Time How long it takes an AI system to process input and generate a response. For interactive use, low latency is crucial—you want near-instant responses to queries. For batch processing, higher latency might be acceptable if it enables more thorough analysis. Latency affects user experience and workflow integration; systems with high latency won't be used effectively even if their results are excellent.

Model Selection / Model Choice Choosing which AI model to use for particular tasks based on their different capabilities, costs, and characteristics. Some models excel at reasoning, others at speed, others at handling long documents. Sophisticated legal AI platforms might use different models for different tasks—one for initial retrieval, another for detailed analysis, another for generating summaries. Understanding model characteristics helps you select the right tool for each job.

Pipeline / AI Pipeline A series of connected AI processing steps where the output of one step feeds into the next. A discovery pipeline might include: document retrieval -> entity extraction -> relationship mapping -> narrative generation. Pipelines automate complex multi-step workflows, ensuring consistency and enabling sophisticated analysis that would be difficult to coordinate manually.

Prompt Template / Prompt Library Pre-written prompt structures that can be filled in with case-specific details and reused across similar tasks. A firm might develop prompt templates for contract analysis, privilege review, or deposition summaries. Templates ensure consistency, codify best practices, and let less experienced users benefit from expert prompt engineering. Building a prompt library is a best practice for legal AI implementation.

Quality Control / Validation Processes for verifying AI outputs against ground truth or expert review. In legal AI, quality control is essential for professional responsibility. This might involve spot-checking AI summaries against source documents, having senior attorneys review AI-assisted issue coding, or comparing AI privilege determinations against human review. Quality control procedures should be documented and consistently applied.

Scalability The ability of an AI system to handle increasing volumes of work without proportional increases in cost or time. Scalable systems can analyze ten documents or ten thousand documents using essentially the same process. For legal work, scalability means you can apply the same AI capabilities to small matters or massive document collections, with costs growing slower than the volume of material.

Ethical and Risk-Related Terms

Bias / AI Bias Systematic errors or unfair outcomes in AI systems resulting from biased training data, flawed algorithms, or problematic deployment. Legal professionals must be aware that AI systems can perpetuate historical biases present in their training data. In discovery, this might mean systems are better at understanding documents from large corporations than from small businesses, or better at analyzing formal legal writing than casual emails. Awareness of potential bias is the first step toward mitigating it.

Data Privacy / Information Security Protecting sensitive information from unauthorized access or disclosure when using AI systems. Legal professionals must understand where client data goes when submitted to AI systems, how it's stored and processed, who might have access to it, and how long it's retained. Different AI deployment models (cloud-based, on-premises, private instance) offer different privacy characteristics appropriate for different sensitivity levels.

Explainability / Interpretability The degree to which humans can understand how an AI system reached particular conclusions. For legal applications, explainability is often essential—courts, clients, and opposing counsel may need to understand how AI contributed to work product. Systems that can explain their reasoning (showing which documents they relied on, which facts supported conclusions) are more suitable for legal work than “black box” systems that cannot explain their outputs.

Professional Responsibility / Ethical Obligations Legal professionals' duties of competence, confidentiality, diligence, and supervision when using AI tools. These traditional obligations don't disappear with AI adoption—they require new competencies. Lawyers must understand AI capabilities and limitations, ensure adequate supervision of AI-assisted work, protect client confidentiality when using AI tools, and avoid over-reliance on AI outputs without independent verification.

Red Teaming / Adversarial Testing Deliberately testing AI systems to identify weaknesses, failure modes, or unexpected behaviors. Before deploying legal AI tools, red teaming might involve trying to make the system hallucinate, testing it on edge cases, or seeing how it handles ambiguous situations. This proactive testing helps identify and address problems before they affect real work product.

Validation / Verification Confirming that AI outputs are accurate and reliable before relying on them. For legal work, validation typically means checking AI-generated content against source documents, verifying that citations are accurate, and ensuring that analysis hasn't missed critical facts or made logical errors. Validation procedures should be proportional to the stakes—higher for work product going to courts or clients, potentially lighter for initial internal analysis.

Emerging and Future Concepts

Artificial General Intelligence (AGI) Hypothetical AI systems with human-level cognitive capabilities across all domains, not just specific tasks. AGI does not currently exist. All present AI systems, including the most advanced LLMs, are narrow AI—capable of impressive performance on specific tasks but lacking general understanding or reasoning capabilities comparable to humans. Legal professionals should be skeptical of claims that current systems have

achieved AGI.

Emergent Capabilities / Emergent Behavior Abilities that appear in large AI models that weren't explicitly trained and weren't present in smaller versions. Researchers have observed that as models grow, they sometimes spontaneously develop capabilities like multi-step reasoning or certain types of mathematical calculation. Understanding that some AI behaviors are emergent rather than designed helps set appropriate expectations—models may sometimes exceed expectations but may also fail in unexpected ways.

Frontier Models The most advanced, capable AI systems available at any given time. As of this writing, frontier models include GPT-4, Claude Sonnet 4, Gemini Ultra, and a few others. Frontier models typically have the largest context windows, strongest reasoning abilities, and most sophisticated language understanding. For complex legal analysis, frontier models generally produce better results, though they may cost more and run slower than smaller models.

Neural Symbolic AI Systems combining neural networks (pattern recognition) with symbolic reasoning (logic and rules). Pure neural networks excel at pattern matching but struggle with logical reasoning; symbolic systems excel at logic but can't handle ambiguity. Hybrid systems might offer the best of both for legal work, where you need both pattern recognition (finding relevant documents) and logical reasoning (applying legal rules). This is an active research area that may significantly improve legal AI capabilities.

Synthetic Data / Synthetic Training Data Artificially generated data used to train or fine-tune AI models. For legal applications with limited training data (rare case types, specialized practice areas), synthetic data generated from expert-crafted examples might enable effective AI systems. However, models trained on synthetic data risk learning artifacts of the generation process rather than real-world patterns, so this technique requires careful validation.

Understanding These Terms in Practice

The terminology of Generative AI can be overwhelming, but you don't need to master every technical detail to use AI effectively in legal practice. Focus on understanding concepts in three categories:

Essential for all users:

- What Generative AI is and how LLMs work
- Prompting and prompt engineering
- Context windows and token limits
- Hallucination and the need for verification
- RAG and how AI grounds responses in your documents
- Professional responsibility and ethical obligations

Important for power users:

- Agents and agentic systems
- Chain-of-thought and few-shot learning
- Temperature and other parameters affecting outputs
- Embeddings and semantic search
- Tool use and function calling

- Quality control and validation procedures

Useful for specialists and strategists:

- Fine-tuning versus prompting tradeoffs
- Multi-agent architectures
- Model selection for different tasks
- Emerging capabilities and frontier models
- Neural symbolic approaches
- Bias, fairness, and interpretability concerns

The field of Generative AI is evolving rapidly. New capabilities emerge regularly, and terminology continues to develop. The most important skill isn't memorizing every definition but developing the ability to learn continuously about new developments and evaluate how they might benefit your legal practice.

When you encounter unfamiliar terms in AI product descriptions, research papers, or vendor presentations, understand that the fundamental concepts in this glossary provide the foundation for understanding more specialized or newer terminology. As long as you grasp core concepts like how LLMs work, what agents can do, why grounding matters, and how to prompt effectively, you have the foundation needed to evaluate and adopt new AI capabilities as they emerge.

The future of legal practice will increasingly depend on thoughtful integration of AI capabilities with human expertise. Understanding these terms isn't about becoming a data scientist—it's about being a sophisticated consumer and user of AI tools, capable of making informed decisions about what technologies to adopt, how to implement them responsibly, and where human judgment remains essential.

Our perspective is this. The question isn't whether AI will transform legal practice, but how quickly and completely legal professionals will adapt to and leverage these powerful new tools. Understanding the language of AI is the first step in that adaptation.



APPENDIX TWO

PROMPT ENGINEERING GUIDE FOR DISCOVERY TASKS

APPENDIX TWO

Prompt Engineering Guide for Discovery Tasks

The shift from keyword search to conversational interaction with Generative AI represents one of the most significant changes in how discovery professionals work with documents. This appendix provides practical guidance for developing effective prompting strategies across different discovery scenarios, from simple questions to complex investigative frameworks.

The Fundamental Shift: From Search to Conversation

Traditional discovery platforms required you to think like a database, translating your information needs into Boolean operators, proximity connectors, and carefully constructed search strings. Generative AI reverses this paradigm. Instead of adapting your thinking to match the technology, the technology adapts to match your natural way of thinking and communicating.

This shift has profound implications for how we approach document review and investigation. When you ask a question of a Generative AI system, you're not simply retrieving documents that match certain keywords. You're engaging in an analytical conversation where the system interprets your intent, synthesizes information across multiple sources, identifies patterns and relationships, and presents findings in the format most useful to your investigation.

The quality of this conversation depends largely on how well you frame your questions and structure your prompts. While natural language makes prompting accessible, understanding the principles of effective prompt construction dramatically improves your results.

Prompting Fundamentals: Starting Simple

If you can ask a question, you can use Generative AI effectively. The most important thing to remember is that natural language works. You don't need special commands or syntax. Type your question as you would ask a colleague, and the system will understand your intent.

Consider these fundamental truths about prompting: Natural language is powerful—full sentences and questions work better than keyword strings. Specificity improves results—adding relevant details helps the system understand exactly what you're seeking. Prompting is iterative—you can refine or follow up on answers to dig deeper without starting over. Each response becomes part of the context for subsequent questions. Conversation beats commands—the more you treat prompting as a dialogue rather than a search query, the better your results will be.

Starting with a simple question is always a reasonable approach. For identifying key people: "Who were the main people involved in approving the project? For each, tell me their title or role and how they were involved." For building timelines: "When did discussions about environmental compliance first begin and how did they evolve? Create a timeline of events in table format with columns for date, person, event summary and document reference." For surfacing problems: "What concerns were raised about the product before it launched? List each con-

cern in a bold heading and then summarize the discussions pro or con about those concerns in bullets beneath your headings.”

These conversation-like prompts yield meaningful answers rather than search results requiring hours of review. As you become more comfortable, you can enhance these simple prompts using more advanced techniques.

Anatomy of an Effective Prompt

While your prompt can begin as a simple question, adding structure to your request produces better results. An effective prompt for discovery tasks typically includes five elements:

Context: Briefly explain the matter and relevant background. Context helps the system understand the significance of different information and prioritize accordingly.

Natural language: Ask questions and make requests in conversational form. The more naturally you communicate, the better the system understands your intent.

Key details: Include specific names, dates, technical terms, or other identifiers relevant to your investigation. These anchor points help focus the analysis.

Specific request: Clearly state what information you’re seeking. Vague prompts produce vague results. Precision in your request yields precision in the response.

Desired output format: Specify how you want information presented—as a narrative summary, a table, bullet points, a timeline, or some other structure that suits your analytical needs.

Let’s examine how these elements work together through progressive refinement of a basic prompt.

From Basic to Structured

Basic Prompt: “What happened with the merger?”

This prompt will likely produce a general response, but it lacks the specificity and context that would make the answer truly useful for discovery purposes.

Adding Context and Details: “This is a securities case involving allegations that XYZ Corporation made misleading statements about its merger with ABC Industries in March 2023. What happened with the merger, particularly focusing on events between January and April 2023 and involving the CEO, CFO, and outside counsel? I need to understand when key decision makers first learned about potential valuation problems and how they discussed those issues internally versus publicly.”

Context immediately focuses the analysis on legally relevant information. Specific timeframes and key players narrow the scope. The explicit information need about knowledge and timing focuses the analysis on elements relevant to your legal theory.

Specifying Output Format: “This is a securities case involving allegations that XYZ Corpora-

tion made misleading statements about its merger with ABC Industries in March 2023. What happened with the merger, particularly focusing on events between January and April 2023 and involving the CEO, CFO, and outside counsel? I need to understand when key decision makers first learned about potential valuation problems and how they discussed those issues internally versus publicly. Create a timeline in table format showing date, person, what they learned, source of their knowledge, and how this compared to public statements.”

The structured output request ensures you receive information in a format that directly supports your analytical needs and can be easily incorporated into work product. Notice how the vague initial question evolves into a focused investigative request that produces actionable analysis rather than raw search results.

Advanced Prompting Techniques

As you become comfortable with basic prompting, several advanced techniques can enhance the quality and relevance of responses.

Multi-Part Prompts

Complex investigations often require multiple pieces of information analyzed together. You can structure prompts to request several related analyses in a single interaction:

“This investigation concerns potential FCPA violations related to payments made through the company’s Dubai subsidiary between 2020-2023. Please analyze the following:

1. Identify all payments over \$10,000 made to government officials or entities classified as state-owned, including date, recipient, amount, stated purpose, and approvers.
2. For each payment identified, determine what due diligence was documented before the payment was made and whether any concerns were raised internally about the legitimacy or appropriateness of the payment.
3. Identify any changes to payment procedures, compliance policies, or approval requirements during this period and what prompted those changes.
4. Summarize any internal communications where employees expressed concerns about the legality or propriety of payment practices in the Dubai office.

Present your findings for each item separately with supporting document references.”

This structured approach ensures comprehensive analysis while maintaining clear organization of results.

Progressive Investigation Framework

Effective investigators often use a sequence of prompts that progressively deepen understanding. Each prompt builds on previous findings, moving from broad assessment to detailed analysis.

Initial assessment: “Provide an overview of all documents relating to the Alpha Project budget overruns. Identify key personnel involved, timeline of significant events, and major areas of concern.” This broad opening prompt gives you the landscape and helps identify what aspects deserve deeper investigation.

Targeted follow-up: “Based on the initial findings, focus specifically on the communications between the Project Manager and Finance Director between March and May 2023. What explanations were given for the cost overruns, and how did the Finance Director respond to each explanation?” Having identified key players and timeframes, you can now drill into specific interactions and relationships.

Pattern identification: “Analyze the explanations given for budget overruns across all projects in 2022-2023. Identify recurring patterns, common justifications, and how the accuracy of initial budget projections compares across different project teams.” Zooming back out with new context allows you to identify systemic issues that might not be apparent from examining a single project.

Synthesis and implications: “Based on all previous analyses, summarize the systemic factors contributing to budget overruns, the effectiveness of existing controls, and potential areas of concern for regulatory compliance or financial reporting.” The final synthesis connects individual findings to broader patterns and legal implications.

Instructing Output Structure

The format in which you receive information significantly impacts its usefulness. Different analytical needs require different presentation formats. You can explicitly instruct the system how to structure its response.

For temporal analysis: “Create a timeline in table format with columns for date, person, event description, significance, and document reference.”

For comparative analysis: “List each public statement in a bold heading, then provide indented bullet points showing how internal discussions aligned with or contradicted that statement.”

For relationship mapping: “For each person identified, create a section with their name as the heading, then use subheadings for ‘Role and Title,’ ‘Key Relationships,’ ‘Areas of Responsibility,’ and ‘Notable Actions or Statements.’”

Clear output instructions ensure you receive information in a format that directly supports your analytical workflow and can be incorporated into work product with minimal reformatting.

Prompting for Common Discovery Tasks

Different types of investigations require different prompting approaches. The following examples demonstrate effective techniques for common discovery needs.

Identifying Admissions and Key Statements

When searching for admissions or acknowledgments, specificity about who might have made statements and what types of acknowledgments you're seeking produces better results.

"Did any company representative acknowledge knowing about the safety issues in the Model X device before the recall announcement on June 15, 2023? For each person who made such an acknowledgment, provide their name, title, role, the exact nature of what they acknowledged knowing, when they claim to have learned it, and from what source."

This prompt identifies not just the existence of admissions but also the context and timing that determines their legal significance.

Mapping Evolving Knowledge and Awareness

Understanding when key players learned critical information and how their understanding evolved is central to many investigations. Timeline prompts that track progression of knowledge are particularly effective.

"When did key personnel first become aware of the accounting irregularities in the European division, and how did their understanding evolve over time? Create a progressive timeline showing when each person learned new information, what specifically they learned, from whom they learned it, and how their characterization of the issue changed in subsequent communications."

This type of prompt surfaces the critical "who knew what when" information that underlies many liability theories.

Comparing Internal and External Communications

Discrepancies between what organizations say publicly and what they discuss internally often form the basis of fraud or misrepresentation claims. Comparison prompts should explicitly structure the analysis around specific public statements.

"How did the company's public statements about the merger compare with internal discussions among executives? List each public statement in a bold heading and then provide indented bullet points showing how internal discussions compared to or contradicted each statement. Focus particularly on revenue projections, expected synergies, and integration timelines."

This prompt produces a structured analysis that maps directly to elements of potential claims and can be used to draft interrogatory responses or deposition outlines.

Identifying Missing Information

One of the most powerful applications of Generative AI in discovery is identifying what should exist but doesn't. These prompts leverage the system's knowledge of standard practices and documentation.

“Based on standard corporate governance practices for public companies, what board-level documentation would typically exist regarding the decision to restate earnings, and what appears to be missing from this collection? Focus particularly on audit committee meetings, discussions with outside auditors, and legal risk assessments.”

Missing document prompts work best when you specify the industry standards, regulatory requirements, or common practices against which gaps should be measured.

Iterative Refinement: The Conversation Continues

The most effective use of Generative AI involves treating each response as part of an ongoing conversation. Your initial prompt establishes the foundation, but follow-up questions refine, expand, or redirect the analysis based on what you learn.

When an initial response identifies something significant, immediately follow up to explore it in detail. If a response indicates the CFO raised concerns about revenue recognition in three emails, ask: “Tell me more about those three emails from the CFO to the controller about revenue recognition. For each email, provide the date, a summary of the specific concerns raised, how the controller responded, whether the concerns were escalated further, and whether any changes were made to revenue recognition practices as a result.”

When a response suggests there might be additional relevant information, expand the scope. If concerns were raised about one product line, ask: “Were similar concerns raised about other product lines during the same period? Compare the types of concerns, who raised them, and how management responded across all product lines.”

When terminology, references, or relationships in a response need clarification, ask immediately. If a response mentions “The Phase 2 approval process was bypassed,” follow up with: “Explain what the Phase 2 approval process normally requires, why it exists, who has authority to bypass it, and whether any documentation exists explaining why it was bypassed in this instance.”

As patterns emerge, use follow-up prompts to test whether they hold across different contexts. If senior managers stopped using email to discuss pricing decisions starting in May 2023, ask: “Did communication patterns about pricing decisions change in other ways starting in May 2023? Look for increased use of instant messaging, phone calls, in-person meetings, or communications through personal devices or accounts. Also identify any events in April or May 2023 that might explain a change in communication practices.”

Technical Considerations and Best Practices

Understanding some technical aspects of how Generative AI systems work helps you craft more effective prompts.

Context Windows and Retrieval

Most modern systems use retrieval-augmented generation (RAG), which means they first search the collection to find relevant documents, then analyze those specific documents to answer your question. This is why specific names, dates, technical terms, and other identifiers

in your prompts matter—they help the retrieval system find the right documents to analyze.

Verification and Quality Control

Generative AI systems can sometimes produce responses that sound authoritative but aren't fully supported by the source documents. Several practices reduce this risk:

Request source citations. Explicitly ask the system to cite specific documents for factual claims: "Provide document names and dates for each assertion."

Ask for quotes. Request relevant quotations from source documents: "Include relevant quotes with document references."

Verify critical facts. For information that will be used in filings, depositions, or other formal settings, always verify against source documents that nothing has been fabricated or mischaracterized.

Use confidence qualifiers. When information is crucial, ask the system to indicate its confidence level: "For each finding, note whether the evidence is clear and direct or whether you're making inferences from indirect evidence."

Common Pitfalls to Avoid

Being too vague: "Tell me about the contract dispute" provides no context, timeframe, parties, or specific information needs. Better: "In the contract dispute between TechCorp and DataSolutions regarding the January 2023 software license agreement, what were the specific performance obligations that TechCorp allegedly failed to meet, what evidence exists regarding whether TechCorp actually met or failed to meet each obligation, and what damages did DataSolutions claim resulted from any failures?"

Thinking in keywords rather than questions: "Merger, fraud, SEC, financial statements, restatement" doesn't give the AI system enough context to understand what analysis or information you need. Better: "Did anyone at the company discuss the possibility that the merger accounting might require restatement or SEC disclosure before the merger was publicly announced? Focus on communications among executives, board members, outside auditors, and legal counsel."

Neglecting to provide context: "What did Johnson know about the safety issues?" Without context about which Johnson, which safety issues, and why timing or sources of knowledge matter, the response cannot focus on what's legally significant. Better: "This is a product liability case involving the AlertX medical device. What did Michael Johnson, the VP of Regulatory Affairs, know about the battery failure issues before the June 2023 FDA submission? For each piece of information he knew, identify when he learned it, from whom, and whether he disclosed it in FDA communications."

Developing Your Prompting Skills

Prompting is a skill that improves with practice and experimentation. Start simple, then elaborate based on results. When you develop prompts that work well for specific types of analysis,

save them—over time you'll build a library of effective prompt templates you can adapt to new matters. Learn from unexpected results by examining why a response wasn't what you expected. Often the disconnect reveals something about how you framed the question or what context you failed to provide. Discovery teams benefit when effective prompts are shared. When someone develops a prompt that produces excellent results, make it available to the team.

From Search to Investigation

The shift from keyword search to conversational prompting represents more than a change in technology. It represents a change in how we think about the investigative process itself. Traditional search required you to hypothesize what terms might appear in relevant documents and craft queries to retrieve those documents for manual review. This approach is reactive and limited by your ability to anticipate what language witnesses and parties used.

Conversational prompting inverts this relationship. Instead of trying to anticipate language, you describe what information you need and let the system find that information regardless of what specific terms were used. Instead of reviewing documents to find answers, you ask questions and receive answers supported by the documents.

This shift enables you to focus on higher-level investigative strategy, legal analysis, and judgment while the system handles the mechanical work of finding and synthesizing relevant information. It allows you to work at the level of meaning rather than the level of words.

The most successful users of Generative AI in discovery view prompting not as a technical skill to master but as an ongoing conversation with their document collection—one that progressively reveals deeper insights and connections that might otherwise remain hidden. Start with a simple question. Let the response guide your next question. Follow the threads that emerge. Build understanding iteratively.

The technology adapts to you. You don't need to learn a new language or master complex syntax. You simply need to ask questions, provide context, specify what you need, and follow where the investigation leads. The future of discovery isn't about better Boolean search. It's about better questions. And if you can ask a question, you can use Generative AI effectively.



— APPENDIX THREE —

ABOUT THE AUTHORS

APPENDIX THREE

About the Authors

John Tredennick, J.D., (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.



William Webber, Ph.D., (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 Search Technologies pioneers the transformation of legal discovery through advanced Generative AI and secure cloud technology. Founded in 2019 by legal technology veterans with over two decades of experience designing and building discovery platforms for the world's largest companies and law firms, Merlin is redefining how legal professionals interact with complex document collections.

Our revolutionary Alchemy platform represents a fundamental shift in legal technology. Rather than forcing professionals to master Boolean operators and construct elaborate keyword searches, Alchemy enables legal teams to ask questions in natural language and receive intelligent, synthesized answers grounded in their specific documents. This transformation—from search results to actual answers—dramatically accelerates investigations, document review, and case analysis while reducing costs and improving accuracy.

The Alchemy Platform: Intelligence, Not Just Search

Alchemy combines fourth-generation Generative AI with traditional search and review capabilities to deliver document intelligence at unprecedented speed and scale. The platform seamlessly analyzes emails, contracts, depositions, medical records, financial documents, and other complex materials, extracting insights and synthesizing findings in seconds rather than days.

What sets Alchemy apart is its integration of multiple AI technologies working together. Advanced semantic search finds conceptually relevant documents even when exact keywords don't match. Large language models analyze and summarize content with nuance and context. Multi-agent systems autonomously break down complex investigations into manageable components, execute searches, evaluate results, and produce comprehensive reports—all while maintaining transparency about sources and reasoning.

The platform excels at tasks that traditionally consumed hundreds of attorney hours: creating detailed deposition summaries, analyzing contracts for specific obligations, mapping relationships between people and transactions, comparing internal communications with public statements, and identifying patterns across thousands of documents. Alchemy doesn't simply retrieve documents for review—it performs the analysis and delivers actionable intelligence.

Security and Privacy by Design

In an era of increasing data breaches and stringent privacy regulations, Merlin's security architecture sets new standards for the legal technology industry. Our unique single-tenant cloud design ensures complete data isolation—each client's documents reside in a separate, dedicated AWS environment with no shared computing resources. This architecture eliminates the data commingling risks inherent in multi-tenant systems and enables compliance with the most demanding global privacy requirements, including GDPR, CCPA, and international data sovereignty mandates.

This security-first approach provides law firms and corporate legal departments the confidence to handle their most sensitive matters on the platform, from white-collar investigations and intellectual property disputes to regulatory matters and high-stakes litigation.



Revolutionary Economics: Cloud Utility Pricing

Merlin introduced the legal industry's first true utility pricing model, allowing clients to reduce hosting costs by up to 60% through pay-by-the-minute cloud infrastructure. Our On/Off Cloud capability enables clients to power down their dedicated environments when not in use, paying only for actual usage time. This innovation not only slashes costs but also supports ESG and green computing initiatives by dramatically reducing energy consumption and carbon footprint.

Traditional discovery platforms require clients to maintain expensive infrastructure 24/7 regardless of actual usage. Merlin's approach aligns costs with work—you pay for the platform when you're using it, not when you're not. For matters with sporadic activity or defined review periods, the savings are substantial.

Recognition and Impact

Alchemy's innovation has earned recognition from industry leaders. ACEDS London named Merlin "Game Changer, Most Innovative Legal Technology for 2024," acknowledging the platform's transformational impact on legal workflows. This recognition reflects what our clients experience daily: the ability to conduct more thorough investigations, identify critical evidence faster, and deliver higher-quality work product at lower cost than traditional approaches allow.

But our real validation comes from the legal professionals who rely on Alchemy for their most important matters. Partners use it to understand complex cases quickly and develop trial strategy. Associates use it to draft comprehensive memoranda in minutes rather than days. Discovery teams use it to analyze massive document collections and surface the needles in the haystack that determine case outcomes.

The Vision: Amplifying Human Expertise

Merlin's mission is not to replace legal professionals but to amplify their capabilities. The platform handles the mechanical work of searching, reading, extracting, and synthesizing information across large document sets, freeing lawyers to focus on judgment, strategy, and advocacy—the uniquely human aspects of legal practice that no AI can replicate.

We envision a future where technology handles routine analytical tasks with speed and consistency, while legal professionals deploy their expertise where it matters most: developing case theories, evaluating risks, advising clients, and advocating in court. This division of labor—AI for scale and speed, humans for judgment and strategy—represents the optimal integration of artificial and human intelligence.

As Generative AI continues to evolve, Merlin remains at the forefront of legal technology innovation, constantly incorporating new capabilities and refining existing ones based on real-world use and client feedback. Our team includes not just software developers but former trial lawyers, data scientists, information retrieval experts, and discovery professionals who understand the practical challenges our clients face and design solutions that address real needs.



Commitment to Responsible AI

Merlin recognizes that with powerful AI capabilities comes responsibility. We build transparency into our systems, ensuring that AI-generated analysis can be verified against source documents. We implement safeguards against hallucination and inaccuracy. We provide clients control over their data and clarity about how AI processes their information. We design our tools to maintain attorney-client privilege and work product protection.

Our platform doesn't make decisions—it provides analysis and recommendations that legal professionals evaluate and act upon. This human-in-the-loop approach ensures that professional judgment remains central to legal work while technology accelerates the analytical processes that inform that judgment.

Join the Transformation

The shift from keyword search to conversational intelligence, from document retrieval to document understanding, from manual review to AI-assisted analysis represents one of the most significant advances in legal technology history. Merlin is leading this transformation, not through hype about what AI might someday do, but through working systems delivering measurable value today.

Whether you're investigating potential fraud, preparing for trial, conducting due diligence, managing litigation, or analyzing contracts, Alchemy provides capabilities that were simply unavailable a few years ago. The question is no longer whether AI will transform legal work—it's whether your organization will lead that transformation or struggle to catch up.

Merlin Search Technologies is headquartered in Denver, Colorado, serving law firms, corporate legal departments, and government agencies worldwide. For more information about how Alchemy can transform your legal workflows, visit www.Merlin.Tech or contact us at info@Merlin.Tech.



GENERATIVE AI

FOR SMART DISCOVERY PROFESSIONALS

4TH EDITION

© 2025 Merlin Search Technologies, Inc. All rights reserved.

This electronic book may be freely shared and distributed for noncommercial purposes only.
You may not use this work for commercial purposes without prior written consent.

This work is licensed under a Creative Commons Attribution-NonCommercial- ShareAlike 4.0 International License. For details,
visit <https://creativecommons.org/licenses/by-nc-sa/4.0/>

About This Edition

The fourth edition reflects both the maturation of Generative AI and our deeper understanding of its practical application in discovery and investigations. Since the third edition, AI systems have gained longer context windows, more accurate outputs, and sophisticated multi-step reasoning capabilities. Most significantly, agentic AI systems can now plan, execute, and refine investigations autonomously—conducting tasks that required extensive human effort just months ago. The legal profession has also gained clearer ethical frameworks from the ABA and state bar associations, providing guidance on responsible AI use in practice.

This edition demonstrates how discovery professionals can use these tools to analyze vast document collections, synthesize testimony across dozens of depositions, identify patterns in complex datasets, and generate detailed investigative reports—all while maintaining quality control and meeting defensibility standards for adversarial proceedings.

What You'll Find in This Book

Part One: Fundamentals of Generative AI provides the technical foundation you need to understand how Large Language Models work, how they're trained, and what makes them different from earlier systems. We address critical questions about data security, hallucinations, and context windows, and explore the ethical obligations that govern AI use in legal practice.

Part Two: Agentic AI and Reasoning Models introduces AI systems that can plan multi-step workflows, use tools strategically, and iterate toward goals. You'll learn the difference between prompting an LLM and deploying an agent, understand single-agent and multi-agent systems, and examine the unique ethical considerations that autonomous systems raise.

Part Three: Practical Applications in Discovery and Investigations demonstrates these concepts through real-world examples, showing how AI analyzes document collections, extracts key information, synthesizes testimony at scale, and even crafts compelling legal arguments. Each example includes specific methodology, results, and critical analysis of strengths, limitations, and defensibility concerns.

Whether you're beginning to explore AI tools or looking to deploy more sophisticated systems, this book provides the foundation and practical guidance you need to work effectively with Generative AI in discovery and investigations.

About the Authors



John Tredennick is CEO and founder of Merlin Search Technologies, which leverages Generative AI for legal discovery and investigations. A former trial lawyer and litigation partner, he founded Catalyst, an international ediscovery technology company acquired in 2019. He has authored eight books on legal technology and served as Chair of the ABA's Law Practice Management Section.



Dr. William Webber is Chief Data Scientist at Merlin Search Technologies. He completed his PhD in Information Retrieval at the University of Melbourne and post-doctoral research at the University of Maryland's E-Discovery Lab. With over 30 peer-reviewed publications in information retrieval and machine learning, he brings deep expertise in AI and statistical measurement to legal technology.

© 2025 Merlin Search Technologies, Inc. All rights reserved.

This electronic book may be freely shared and distributed for noncommercial purposes only.
You may not use this work for commercial purposes without prior written consent.

This work is licensed under a Creative Commons Attribution-NonCommercial- ShareAlike 4.0 International License. For details, visit <https://creativecommons.org/licenses/by-nc-sa/4.0/>