# The Leery Lawyer's Guide to AI and LLMs in Trial Practice

Updated 2026

Craig Ball

**The Leery Lawyer's Guide to AI and LLMs in Trial Practice**

**By Craig Ball**


## Contents

## Introduction

This paper is a wholesale update to my 2025 primer, *Practical Uses for AI and LLMs in Trial Practice*. It remains aimed at tech-wary, veteran trial lawyers, but it has been broadened to serve those who have cautiously woven AI tools into their litigation practice. Since large language models like ChatGPT burst onto the scene three years ago, the hype and fearmongering haven't let up. Every day brings fresh alarm: ETHICAL PITFALLS! DEEP FAKES! JOB LOSS! COPYRIGHT INFRINGEMENT!

The noise has only grown louder. In 2025–26, the legal press ran out of synonyms for *transformational*, and vendors now insist that every litigation tool has "AI inside," whether or not that claim would survive a Daubert hearing. Still, beneath the marketing froth lies something real and durable: AI has become a routine—if imperfect—instrument in modern trial practice. Nothing in my forty-odd years at the Bar has had such a swift and profound effect on the way lawyers work.

Some lawyers point to AI's occasional tendency to "hallucinate"—to fabricate information—as an excuse to avoid adoption altogether. Yet we all use AI daily, even if only by reading the summaries atop Google searches or skimming Amazon reviews. What drives the relentless drumbeat of AI, AI, AI isn't simply the integration of smarter agents into our software (so-called *agentic AI*); it is the paradigm-shifting species of generative AI known as *large language models*, or LLMs—an initialism that no longer denotes a Master of Laws, but the most consequential development in technology since the Internet.

LLMs can talk, read, write, and draw pictures. They can speak in a thousand voices—including yours—dishing out sense and nonsense in equal measure. Their uncanny ability to "sound" human and to string words together in ways that convincingly mimic (and sometimes exceed) human language skills is at once exciting and unsettling.

Despite high-profile instances of lawyers getting into trouble by using LLMs carelessly, the case for employing AI in trial practice today is compelling—even if only as a time-saver for document generation and summarization, tasks where AI's capabilities are undeniable. The economics are now clearer. Sensibly implemented AI reduces drafting time, accelerates motion practice, sharpens discovery management, and improves comprehension of large records. Used prudently, it functions as the world's fastest junior associate—minus the hourly rate and parking validation. Used unwisely, it can embarrass you or worse.

The earliest iteration of this guide grew out of the Litigation Section of the State Bar of Texas devoting the Winter 2024 issue of *The Advocate* to artificial intelligence. Every article was wise and well written—several penned by close friends—but none was *practical* in the sense of helping lawyers use AI in their daily work. That struck me as an unmet need.

As I looked around, I found no articles aimed at guiding trial lawyers who wanted to use LLMs safely, strategically, and without turning their practice upside down. Here, I've tried to fill that gap

by highlighting important considerations and illustrating applications you can try now with minimal expertise, anxiety, or expense. LLMs won't replace your professional judgment, but they can frame issues, suggest language, and translate complex doctrines into plain English. In truth, they can do just about anything that mastery of facts and language can achieve—so long as you remain firmly in the driver's seat.

**But Know This…**

LLMs are unlike any technology tool you've used before. Most of the digital tools in our lives are characterized by consistency: you put the same things in, and the same things come out, in a rigid and replicable fashion. Not so with LLMs. Ask an LLM the same question multiple times, and you are likely to receive somewhat different answers each time. That takes some getting used to.

There is also no single "right" way to interrogate an LLM to be assured of an optimal result. There is no strict programming language or magic set of keywords calculated to achieve your goal. Instead, there are myriad ways to elicit useful information, and—unlike the inflexible and unforgiving tech tools of the past—the most effective approach is often the simplest: interact with your AI tool in a natural, conversational manner, refining your requests as you go.

Throughout this paper, I reference both **ChatGPT** and **Harvey**. ChatGPT now exists in both consumer and enterprise deployments. Harvey is purpose-built for law firms, offering matter-centric workspaces, access controls, retention policies, and auditable logs—features designed to satisfy skeptical general counsel and their insurers. Both platforms are useful. Both are powerful. And both require adult supervision.

## Marketplace Models and Versions

Fueled by a trillion+ dollars in corporate investment, the Generative AI market didn't cool in 2025; it matured. LLMs diversified into two main channels:

1. **consumer/general-purpose models** such as OpenAI's GPT series and Anthropic Claude, and
2. **enterprise-governed systems** like Harvey, deployed behind stronger privacy and audit controls.

Here are current market leaders, with no claim that the list is exhaustive or permanent—because it isn't:

- OpenAI: Offers the GPT series, including GPT-5.2 and multi-modal models capable of ingesting text, images, spreadsheets, and transcripts. These models are accessible through products like ChatGPT and are integrated into various applications.

- Google DeepMind: Developed the Gemini series (successor to Bard), with versions like Gemini 3 Flash and Gemini 3 Deep Think. These models power AI capabilities in Google's services.
- Anthropic: Provides the Claude series, including Claude Opus 4.5 designed for conversational AI applications.
- Meta AI: Developed the Llama series, with Llama 4 Scout and Maverick being notable versions. These models support multimodal inputs (i.e., text and images).
- Microsoft: Introduced the Phi series, such as Phi-4 (that understands text, images and audio) and Phi-4-Mini, marketed as a "small language model" for efficient AI tasks.
- Harvey AI: Built expressly for law firms, Harvey operates within firm-controlled environments, supporting matter-level governance, logging, retention limits, and privacy controls appropriate for supervised legal work.

This article will focus on OpenAI's ChatGPT, specifically its ChatGPT-5 product. I don't feature it because it's necessarily the best or brightest; it's the LLM tool most people use—and increasingly its enterprise deployments reduce privacy anxiety. The current $20 per month subscription tier remains surprisingly affordable compared to billable time squandered staring at a blank pleading page.

I've also expanded my focus to the NOT inexpensive Harvey AI, which you might think of as "enterprise ChatGPT with governance." If you're a solo or small-firm lawyer on a budget, think "ChatGPT—used prudently—can still be safe and effective."

Whatever LLM tool you choose, remember this: *There's a huge difference in the capabilities and user experience between the free versions and the subscription offerings.* ***If you're going to integrate an LLM into your practice, even just to kick the tires, opt for the pay-to-play versions.*** The paid tier unlocks document upload, longer context windows, and improved accuracy. The enhanced performance of GPT-5.2 justifies the modest investment, particularly when compared to the hourly cost of professional staff or consultants.

So, yes, doing the things I describe here will cost you twenty bucks, but you can try it for a month or two, then cancel if you don't see value in continuing to subscribe. I *guarantee* you will see value, if only from the ability to upload content to the tool for analysis and editing.

## Getting Started

Getting started with ChatGPT is surprisingly simple, even for lawyers who aren't tech-savvy but possess intellectual curiosity and critical thinking skills honed through years of trial practice. At its core, ChatGPT is a conversational artificial intelligence tool designed to assist with tasks such as drafting, research, and document analysis. To begin, visit chat.openai.com and create an

account using your email address and date of birth. Upgrade your account to a $20.00 per month **ChatGPT Plus** subscription.

**Harvey Users:** Access occurs inside your firm's workspace. Treat it like a secure review platform, not a personal app. Matter access is controlled, and work is logged—assume someone responsible is watching, because they should be.

Once logged in, you can enter **prompts**—questions or instructions—to generate **responses**. Think of ChatGPT as a hyper-efficient research assistant who can respond instantly to requests for drafting demand letters, summarizing testimony, or brainstorming trial strategies.

To maximize ChatGPT's usefulness, *clarity and specificity* in prompts are key. The tool thrives on detailed instructions.

For example, instead of saying,

"*Draft a request for production for a baby powder cancer case*,"

You'll get a superior result by uploading your complaint to the tool and prompting, *e.g.,*

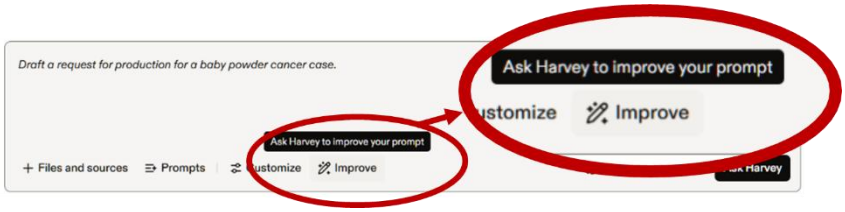
> *"You are plaintiff's counsel in a wrongful death case in a Harris County, Texas civil district court. Draft 25 numbered Requests for Production to defendant Johnson & Johnson regarding alleged cervical cancer caused by exposure to talcum powder containing asbestos.*
>
> *Drafting requirements:*
>
> - *Use Texas practice and terminology; track Texas Rules of Civil Procedure (e.g., 192/196) and proportionality.*
> - *Include Texas-style Definitions and Instructions (parties, "document," "ESI," date range, custodians, form of production).*
> - *ESI specifics: identify relevant systems/custodians; request native or near-native production with agreed metadata; de-duplication and email threading; reasonably usable format; index/cover letter.*
> - *Avoid overbreadth; tie each request to issues, custodians, systems, and time period."*

**Why the detailed prompt is superior**

- **Clarity about role and forum.** By specifying that the LLM is to perform as "plaintiff's counsel" in a Harris County, Texas civil district court, the model defaults to lawyerly language in a plaintiff-side posture and Texas terminology, instead of mixing federal/state or other jurisdictions.
- **Concrete facts and theory**. Naming Johnson & Johnson, cervical cancer, talc, and asbestos anchors the requests to issues you actually intend to discover (design, warnings, testing, adverse events, marketing, regulatory knowledge), not generic product-liability boilerplate.

- **Scope and deliverable defined.** "Draft 25 numbered RFPs" tells the model the format and volume, so you get a serviceable, complete set—not a handful of examples that require rebuilding.
- **ESI and proportionality baked in.** Calling out "include the language needed to obtain electronically stored information" prompts the model to add definitions, form-of-production, metadata, deduplication/threading, and usability terms that make your set meet-and-confer-ready and defensible under Texas proportionality standards.
- **Hallucination control.** Specific party, product, injury, and venue reduce the model's temptation to invent authorities or wander into non-Texas practice. Constrained prompts produce fewer "creative" but unusable requests.
- **Editing burden drops.** When the model knows the audience, issues, and output format, you spend time doing substantive tailoring instead of structural clean-up.
- **Ethical and practical defensibility.** A prompt that elicits targeted, proportional, ESI-aware requests makes your paper trail stronger when opposing counsel (or the court) asks why your terms or formats are reasonable.

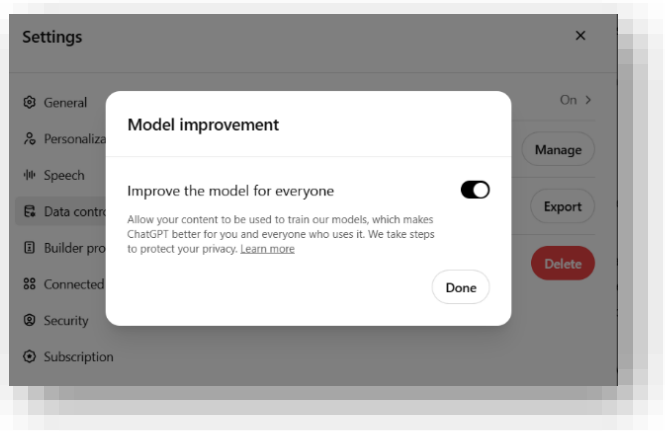
One of the simplest and most effective ways to improve your prompts is by asking an LLM to do it for you. I'll often give ChatGPT a prompt and then ask it to streamline and improve my prompt before I run it. Harvey makes that easier by supplying an "Improve" button dedicated to prompt refinement (right).




## Privacy

As lawyers, we're entrusted with sensitive, privileged, and private information we are obliged to protect. That obligation does not stop at the courthouse door—or at the login screen of an AI platform.

If you're concerned that interactions with ChatGPT may expose confidential information by being used to improve the model, you should adjust the platform's privacy controls before using it for legal work.

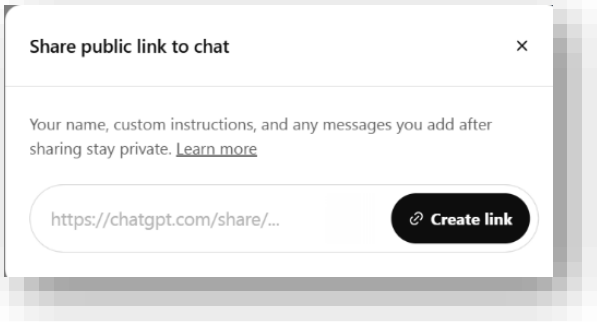



**To modify ChatGPT's privacy settings (consumer and Plus plans):**

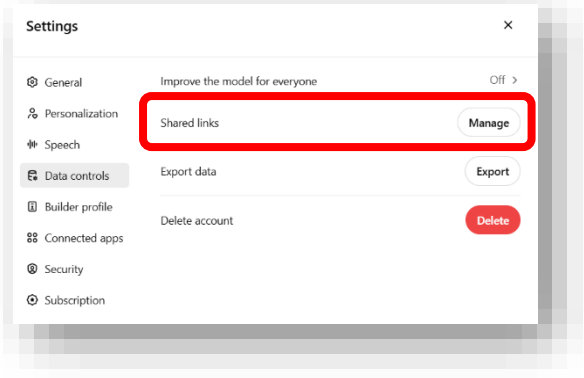
1. Open the ChatGPT app or website.
2. Go to **Settings → Data Controls**.
3. Toggle OFF **"Improve the model for everyone."**

4. Confirm your changes.

Disabling this setting prevents your chats from being used for model training, but it does **not** make the conversation ephemeral or invisible. Chats may still be retained for operational, safety, or legal reasons, and anyone with access to your account can see them.




ChatGPT also allows users to generate **shared links** to conversations. Anyone with the link can view the conversation exactly as it existed when the link was created. These links are convenient—and dangerous. They can be forwarded, indexed, or accessed outside your control. If you share them at all, share them sparingly, only with authorized recipients, and periodically review or revoke them under **Settings → Data Controls → Shared Links → Manage**. As a rule, do not use shared links for privileged or client-confidential material.




Protecting sensitive data does not end with flipping a toggle. It is a necessary step, but no substitute for a layered approach that includes judgment, restraint, and supervision. When in doubt, abstract facts, anonymize parties, or describe issues at a level that conveys the problem without exposing the client.

**Harvey (enterprise law-firm deployments)**

Harvey is designed for legal practice and offers a substantially different security posture from consumer ChatGPT. Work occurs inside **matter-centric workspaces** governed by firm controls. Inputs are not used to train public models, access is role-based, and activity is logged—features that matter to general counsel, insurers, and courts.

That said, Harvey does not eliminate risk. Security depends on **configuration and discipline**:

- Limit workspace access to lawyers and staff assigned to the matter.
- Treat uploads as you would a shared litigation repository: only what belongs in the matter goes in.
- Assume all work is auditable and supervised. That is a feature, not a flaw.
- Do not rely on the platform to decide what is privileged; that judgment remains yours.

Whether using ChatGPT or Harvey, the governing principle is the same: AI is a powerful non-lawyer assistant. You may delegate tasks, but not responsibility. Confidentiality, privilege, and professional judgment remain non-delegable duties.

## Hallucinations

Working with LLMs is like teaching a teenager to drive. Sure, teens have better vision, hearing, and reflexes; but, at any moment, you may find yourself screaming, *RED LIGHT! RED LIGHT!* ChatGPT is merely a tool, one only as capable and trustworthy as your ability to direct and monitor its use. Like any power tool, when you fail to respect its risks and limitations, it can hurt you; operated skillfully, it will make your work easier.

ChatGPT's tendency to hallucinate—*i.e.*, to fabricate information or confidently fill gaps it does not actually know—requires constant vigilance. **ALWAYS cross-check its output**, particularly when legal accuracy is critical. If the tool makes up a case, statute, or quotation, *you* are the one who will have to answer for it. I cannot emphasize this enough. Courts do not want to hear that you failed to verify your citations because "the AI said so."

In practice, hallucinations most often arise when the model is asked to operate beyond the record supplied or is pressed for conclusions where the source materials are silent or incomplete. You can substantially reduce this risk by **instructing the model how to behave**, for example:

- "Use only the materials I provide; do not rely on outside knowledge."
- "If the record is silent or ambiguous, say so."
- "End with a list of assertions that require human verification."

These instructions do not eliminate hallucinations, but they make them easier to spot—and easier to defend against. See also, "Practice Tip: Verification Checklist for Use in Prompts" on the following page.

Harvey reduces hallucination risk by constraining the model to a defined matter and document set, but it does not eliminate it. Even when analysis is limited to uploaded materials, the model can still infer incorrectly, over-generalize, or misread context. Whether you use ChatGPT or Harvey, hallucinations remain a supervision problem, not a software bug.

**Practice Tip: Verification Checklist for Use in Prompts**

*To guard against harm from hallucinations, include the following instruction at the end of any prompt involving legal authority, factual assertions, or record-based analysis:*

**Verification Checklist:**
Before finalizing your response:

1. Identify every legal rule, case, statute, or procedural requirement referenced and flag each for independent attorney verification.
2. Identify every factual assertion not directly quoted from the supplied materials and flag it for confirmation.
3. For summaries of testimony or documents, include page and line numbers or document IDs for all quoted or paraphrased material.
4. If the record is silent, incomplete, or ambiguous on any point, state that explicitly rather than inferring an answer.
5. Do not fabricate citations, quotations, or facts. If you cannot confirm something from the provided materials, say so.

*The goal is not to eliminate lawyer review, but to make review faster, more reliable, and defensible.*

## Customizing Settings for Litigation Efficiency

Modern LLM platforms offer customization features that can materially improve efficiency in litigation and trial work, but they operate differently across tools.

**ChatGPT**

- **Response Style (formerly "Temperature"):**
  While earlier versions exposed a numerical "temperature" setting, most current ChatGPT interfaces control precision through prompt language rather than sliders. For litigation tasks—discovery requests, deposition outlines, motion drafts—explicitly instruct the model to be *formal, precise, and conservative*. Reserve creative latitude for brainstorming themes, demonstratives, or voir dire framing.

- **Custom Instructions:**
  ChatGPT allows persistent instructions that shape behavior across sessions. For example: *"Assume I am a Texas trial lawyer. Use plain English. Do not invent authority. Flag any statement requiring verification."*
  These instructions reduce rework and reinforce disciplined outputs. They are set at **Settings → Personalization → Custom Instructions**.

- **Memory:**
  ChatGPT's memory feature can retain preferences or context across chats. This can be helpful, but it also poses risks in legal work, including cross-matter contamination and inadvertent retention of sensitive information. Review and manage memory regularly at **Settings → Personalization → Memory → Manage**, and consider disabling it entirely for client-specific work.

**Harvey**

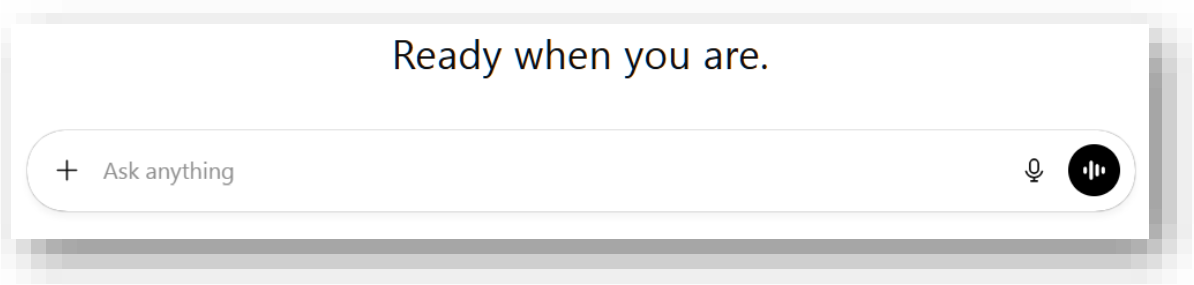
Harvey does not rely on temperature sliders or long-term conversational memory. Instead, its efficiency controls are structural:

- **Matter-centric scope:** Analysis is confined to the documents and prompts within a specific matter.
- **Document constraints:** Outputs can be limited to the uploaded record, reducing drift.
- **Auditability:** Prompts and outputs are logged, reinforcing disciplined use.

These features make Harvey better suited for sustained case work, but they do not replace the need for careful prompt design and lawyer review. Precision still comes from how you frame the task, not from the platform alone.

## The ChatGPT User Interface and its Token, Upload, and Context Limits

When users interact with ChatGPT, they typically do so through the **Chat Window**, a text box with icons in the lower left corner. Users enter prompts and receive responses in the Chat Window. The plus sign initiates a file upload (and more).[1]




If you've experimented with ChatGPT, chances are you've been frustrated by limitations stemming from a failure to understand the tool's **token, upload, and context limits**, even if you didn't appreciate *why* things weren't working.

[1] That black circle with a waveform on the right of the Chat Window opens Voice Mode, allowing you to converse with ChatGPT as you would with the world's best-informed colleague who is unfailingly polite and helpful. Prepare to be awe-stricken.

**Tokens** are units of text used to measure both input and output. Tokens can be as small as a single character or as large as a short word, depending on structure. Every prompt, response, space, and punctuation mark consumes tokens, and longer or compound words consume multiple tokens.

As a rule of thumb, one page of text equals roughly **375–400 tokens**, and **10,000 tokens ≈ 7,500 words ≈ 25 pages**. That estimate remains useful even as platform limits change.

Think of the **context window** as the size of a workspace where the model can "hold" a finite amount of information at one time. The **upload limit** is more like the size of an entire library. You may upload a very large document, but only a portion of it can be actively considered at once. The model can retrieve and move portions of the document into the workspace as needed, but it cannot reason over the entire library simultaneously.

This distinction explains why large uploads sometimes appear to "stall," truncate, or produce incomplete answers. The document may be present, but the relevant portion may not be in the active workspace when the model responds.

As I write this in early-2026, ChatGPT's paid tiers allow users to upload large files and work with extended context windows. However, the precise limits vary by model and plan and change often. In **ChatGPT-5.2** (single-user monthly subscription) these limitations are substantially expanded but still very real:

- ### Token Limit (per chat/working context):
  ChatGPT 5.2 supports a **very large context window—on the order of hundreds of thousands of tokens per chat window**, encompassing both prompts and responses. That translates into *hundreds of pages* of text under ideal conditions—far beyond what was possible a year ago. But this is a working limit, not a guarantee that everything supplied will be considered simultaneously.

- ### Upload Limits:
  Users may upload **multiple large documents per chat** (commonly on the order of a few dozen), with individual file size limits typically in the **hundreds of megabytes** for text-based documents and lower limits for spreadsheets and images. Upload capacity is best understood as the size of the *library*, not the workspace. Uploading a very large file does not mean the model will reason over it all at once. Instead, ChatGPT retrieves portions of uploaded files as needed, subject to the context window. In practice, very large or complex files may process slowly, partially, or fail altogether—particularly when multiple large files are uploaded at once.

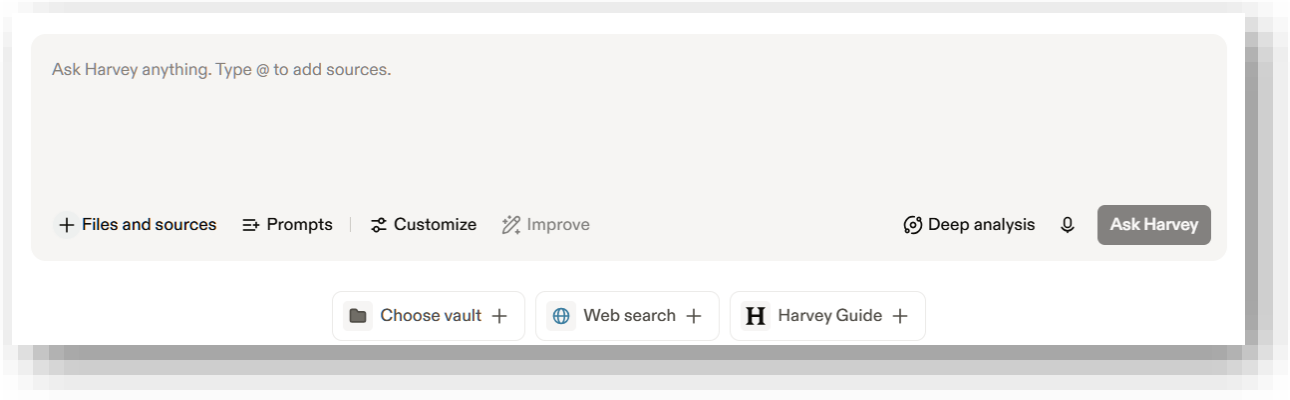
- ### Context Limit:
  The context window—the amount of information the model can actively "hold in mind" during a single interaction—is now dramatically larger than in early 2025, but it remains finite. As conversations grow, earlier prompts and responses are displaced by newer

material. Even uploaded documents may fall outside the active context unless key portions are summarized or reintroduced.

Is this confusing? You bet! Think of the context window as the size of a workspace where you can spread out a very large—but still limited—set of papers at one time. The file upload limit is more like the size of the entire library. You can upload a large book (or several), but only part of it can be open on the workspace at once. The expanded upload and context limits in ChatGPT 5.2 make incremental and sectional analysis far more practical than before, but they do not eliminate the need to break large records into manageable pieces.

## The Harvey User Interface and its Token, Upload, and Context Limits

When users interact with **Harvey**, they do so through a conversational workspace anchored to a **specific vault or matter**, rather than a general-purpose chat window. At the top of the interface is a single prompt field ("Ask Harvey anything"), where users enter questions or instructions.




Unlike consumer chat tools, this field is tightly integrated with sources and controls that govern what Harvey may see and use.

Directly beneath the prompt field are controls that shape the scope and behavior of the interaction. The **"Files and sources"** option allows users to attach documents or select existing materials already stored within the chosen vault. The **"Choose vault"** control determines the matter or repository to which the interaction is confined—a critical distinction, as Harvey's analysis is limited to the documents and sources available within that vault.

Additional interface controls—such as **"Prompts," "Customize,"** and **"Improve"**—allow users to reuse structured prompts, adjust output preferences, or refine prior responses. The **"Web search"** and **"Harvey Guide"** options signal whether the model may consult external reference material or remain confined to internal sources. When enabled, **"Deep analysis"** instructs Harvey to spend more time reasoning across the available materials, though it does not alter the underlying limits on how much information can be considered at once. The interaction is executed by selecting **"Ask Harvey."**

If you've experimented with Harvey, you may still encounter frustrations traceable to a misunderstanding of the platform's **token, upload, and context limits**, even though those limits are mediated through matter-based controls rather than exposed as raw numbers.

As with other large language models, **tokens** remain the unit used to measure both input and output. Every prompt, response, space, and punctuation mark consumes tokens, and longer or compound words consume multiple tokens. The same practical rule of thumb applies: roughly **375–400 tokens per page** of text. This estimate remains useful when deciding how much material to analyze in a single request.

As of early-2026, Harvey supports the ingestion of substantial document sets within a vault and provides extended working contexts compared to consumer chat tools. However, these capabilities do not eliminate fundamental constraints. The precise limits depend on firm configuration and deployment and may change over time. As with ChatGPT, understanding how Harvey *operates* is more valuable than memorizing numerical ceilings.

Think of a Harvey **vault** as a curated library tied to a legal matter. The model has access to the contents of that library, but its **active reasoning context**—the workspace in which it can "hold" and analyze information at one time—is necessarily smaller. Harvey can retrieve documents from the vault as needed, but it cannot reason across the entire contents of the vault simultaneously.

This distinction explains why Harvey may sometimes return incomplete, overly narrow, or seemingly inattentive answers even when the relevant document exists within the vault. The document may be present and available, but it may not have been drawn into the active context at the moment the response is generated. Effective use of Harvey therefore depends on directing the model's attention—by selecting appropriate sources, constraining prompts to defined subsets of material, and building analysis incrementally rather than attempting to process an entire matter in a single pass.

## Why Tokens Matter—and How Uploading Files Actually Works

**ChatGPT**

These limits mean that **ChatGPT may fail to upload files that are too large** or may truncate longer documents such that ingestion simply peters out when upload or token limits are reached. Even when a document uploads successfully, ChatGPT may process only portions of it at any given time.

As a conversation grows, ChatGPT will eventually **"forget" earlier prompts, responses, and instructions** as the context window fills and older material is displaced by newer exchanges. This is one reason refining answers late in a long session can be frustrating: the background you carefully established may no longer be "in view." Preserving continuity often requires summarizing key points and reintroducing them in later prompts so they remain within the active context.

In short, **LLMs have strict working-memory limits**. Do not expect ChatGPT to digest or reason coherently across an entire case file in a single interaction.

That reality also shapes how file uploads work. In current versions of ChatGPT, documents are added using the **plus sign ("+") control located alongside the prompt field**. Selecting the plus sign allows users to attach files from their local system or, depending on plan and configuration, from connected repositories. Once attached, those files become available for analysis within that conversation.

Uploading documents enables ChatGPT to perform a wide range of useful tasks: summarizing deposition transcripts, analyzing medical records, extracting timelines, reviewing spreadsheets, and drafting work product informed by the uploaded materials. ChatGPT works best with **Word documents, Excel spreadsheets, PDFs containing selectable text, and CSV files** exported from databases. Scanned or image-only PDFs remain less reliable unless they include usable text layers.

Attaching a document, however, does not mean the model has fully absorbed it or will continuously remember it. Large documents may be processed only in part, and details may fall out of context as the conversation progresses. Effective use requires **explicit direction**—for example, telling the model which sections to focus on—and often follow-up prompts that pull additional portions of the same document into the active workspace.

The upshot is that the dream of uploading every deposition, exhibit, and production in a case and letting ChatGPT "figure it out" is not yet practical. But within these limits, ChatGPT can still manage minor miracles—so long as you are willing to **break work into stages, process long documents in sections, and divide large datasets into manageable chunks**.

**Harvey**

Harvey changes the container, but not the underlying physics. Documents are added through the **"Files and sources"** control beneath the prompt field and are ingested into a selected **vault or matter**, where they persist across sessions. This persistence can make it feel as though Harvey "knows" the entire case.

In reality, Harvey remains subject to the same fundamental constraints as other large language models. Although documents reside within the vault, any single response is generated within a bounded **active analysis context**. Harvey can retrieve documents from the vault as needed, but it cannot reason across the entire contents of a large matter simultaneously.

This explains why Harvey may sometimes return incomplete or narrowly focused answers even when the relevant document exists within the vault. The document may be present and indexed, but it may not have been drawn into the active context at the moment the response was generated. As with ChatGPT, effective use depends on **directing attention**—by identifying specific documents, narrowing the scope of analysis, or building conclusions incrementally across multiple prompts.

Harvey's matter-centric structure reduces organizational friction and improves governance, but it does not eliminate cognitive limits. Depositions are best summarized one at a time. Expert reports benefit from section-by-section analysis. Chronologies are most reliable when built iteratively. In both platforms, the practical lesson is the same: **design your workflow to fit the size of the workspace, not the size of the library.**

**Practice Tips (Avoiding the Most Common Missteps re: Context Windows)**

**Upload ≠ analyze.** Attaching a document doesn't mean the model has read or reasoned over all of it.

**Expect forgetting.** As conversations grow, prompts and instructions fall out of context. Restate key instructions.

**Name your sources.** Specify which documents or excerpts to use—particularly in Harvey's vault environment.

**Work incrementally.** If answers seem incomplete, narrow the task and rebuild rather than starting over.

**Think like a trial lawyer!** You don't try an entire case in one sitting. Don't ask an AI to do so.

## Chunking Large Uploads

"Chunking" is the practice of breaking large uploads into pieces small enough for an AI tool to handle accurately and reliably. Because large language models can only "hold" a limited amount of material at one time, effective use depends less on uploading everything and more on deciding what to present, when, and in what order. Chunking mirrors how trial lawyers already work: we don't absorb an entire case file in one sitting, and we don't try every issue at once. By dividing records, transcripts, and productions into logical segments—and building conclusions incrementally—we improve accuracy, reduce omissions, and retain control over the analysis. The playbook below offers concrete ways to chunk common litigation materials so the AI works within its limits instead of stumbling over them.

### Chunking Playbook (With Sample Prompts)

**Depositions**

Work one witness at a time. Begin with a high-level summary, then process longer transcripts in logical segments (liability, damages, experts). Synthesize only after the parts are complete.

*Sample prompt:*

*You are a Texas trial lawyer. Using only the attached deposition transcript, prepare a concise narrative summary of the witness's testimony on liability. Cite page and line numbers. If the transcript is silent or ambiguous on any point, say so. End with a Verification Checklist.*

**Expert Reports**
Divide expert reports into discrete sections—background, methodology, assumptions, opinions, and conclusions. Analyze each section separately before requesting a consolidated critique or cross-examination outline.

*Sample prompt:*
*Using only the attached expert report, analyze the methodology section. Identify stated assumptions, potential weaknesses, and areas for cross-examination. Quote or cite to the report where possible. Do not infer facts not stated. End with a Verification Checklist.*

**Productions and Document Sets**
Start with targeted subsets rather than entire productions. Focus on key custodians, date ranges, or document families, then expand outward as needed.

*Sample prompt:*
*Review the attached documents from Custodian A dated January–June 2025. Summarize recurring themes and identify documents likely relevant to contract formation. List document identifiers. If relevance is uncertain, flag it for review.*

**Medical Records**
Chunk medical records by provider or time period. Analyze each segment for chronology, gaps, and red flags before combining results.

*Sample prompt:*
*Using only the attached medical records from Provider X (2019–2021), prepare a treatment chronology noting dates, diagnoses, and procedures. Identify gaps in care or inconsistencies. Cite record pages. End with a Verification Checklist.*

**Chronologies and Timelines**
Build chronologies incrementally, confirming accuracy at each stage before extending the timeline.

*Sample prompt:*
*Based solely on the attached documents covering March–May 2025, prepare a dated chronology of events relevant to notice and response. Cite each entry to document and page. Do not speculate beyond the record.*

**Large Spreadsheets and Databases**
Process data in logical slices—by date range, category, or variable set—before seeking broader conclusions.

*Sample prompt:*

*Review the attached spreadsheet for transactions from Q1 2025. Identify unusual patterns, outliers, or missing fields. Describe findings in plain English and note any assumptions requiring verification.*

## Practical Applications for Litigation and Trial Practice

Large language models are most useful in litigation when they perform tasks lawyers already do—just faster and with fewer false starts. Used properly, they assist with framing, organization, and first-pass drafting, while leaving judgment and verification where they belong: with counsel.

### Drafting Discovery Requests

AI tools excel at translating narrative descriptions into actionable discovery requests. You can describe the dispute, the issues in play, and the information you seek, and the model can generate targeted requests tailored to those needs. This is particularly effective for drafting requests for production, interrogatories, and ESI-specific requests that reflect proportionality and scope.

In ChatGPT, this works best when you provide context and constraints in the prompt. In Harvey, it is most effective when requests are generated within a matter using defined custodians, systems, and date ranges already present in the vault.

### Summarizing Voluminous Data Sets

LLMs can sift through testimony, expert reports, or document subsets to produce summaries, chronologies, or issue-focused digests. Rather than reading everything end-to-end, lawyers can direct the model to surface themes, flag anomalies, or extract testimony on discrete topics.
This is especially valuable for depositions, medical records, and expert materials. The key is scoping: summaries are most reliable when limited to defined witnesses, documents, or time periods, and when citations back to the source are required.

### Structuring Information

ChatGPT and Harvey are adept at converting unstructured narrative into **tables, timelines, matrices, and outlines**. This includes witness charts, issue grids, chronology tables, and comparison charts for contracts or expert opinions.

Structured outputs are often more useful than prose. Asking for a table or outline forces clarity, exposes gaps in the record, and creates work product that can be refined or repurposed for motions, mediation, or trial preparation.

### Search Term Generation and Query Assessment

Turning plain-English descriptions into search terms or Boolean queries for discovery is another area where LLMs perform well. They can analyze relevant documents and propose keyword lists,

proximity connectors, and alternative formulations based on how people actually communicate rather than how lawyers guess they might.

LLMs can also critique existing search terms, identifying overbreadth, redundancy, or likely blind spots. This is useful both for improving your own searches and for evaluating the adequacy of an opponent's proposed terms.

### Drafting Pleadings and Arguments

AI tools can assist in drafting and organizing pleadings, motions, and briefs by suggesting structure, sequencing arguments, and refining language. They are particularly useful for first drafts, issue outlines, and transitions.

They are less reliable for unsupported legal authority and should never be trusted to supply citations without verification. Used properly, they help you get from a blank page to a workable draft more quickly, leaving you free to focus on substance and strategy.

### Demonstrative Evidence and Trial Themes

LLMs can assist in **conceptualizing and planning demonstrative evidence**, but their usefulness—and their limits—vary by platform. Neither ChatGPT nor Harvey will produce trial-ready, admissible exhibits. What they can do, if properly directed, is help lawyers think through how complex information might be **organized, sequenced, and explained**.

ChatGPT, particularly in its consumer and enterprise forms, can generate **visuals, diagrams, and mock-ups**, or describe how a demonstrative might look and function. While some might be extraordinary, these outputs are best treated as sketches or talking points rather than finished products. They are helpful for brainstorming timelines, flowcharts, or comparative graphics, but may require translation into proper demonstratives by counsel or a graphics professional.

Harvey, by contrast, is not designed to generate images or visual mock-ups. Its strength lies in **structuring the underlying content** that demonstratives rely on—chronologies, tables, issue matrices, and witness summaries—within the confines of a specific matter. Harvey can help ensure that demonstrative concepts remain grounded in the record and consistent across documents, but the visual execution occurs elsewhere.

In both platforms, the value lies upstream. LLMs can help identify what should be shown, in what order, and for what purpose. The judgment about how—and whether—to show it to a jury remains firmly in the lawyer's hands.

They are also effective at stress-testing trial themes—identifying weaknesses, alternative framings, or juror-friendly explanations that can be refined by counsel.

### Deposition Preparation and Impeachment

LLMs are particularly effective at preparing topic-organized deposition outlines, identifying internal inconsistencies, and surfacing potential impeachment material across multiple

transcripts or documents. When constrained to the record, they can generate contradiction tables keyed to page-and-line citations, helping lawyers see where testimony shifts over time.

### Meet-and-Confer and Proportionality Advocacy

AI tools can assist in drafting meet-and-confer correspondence that explains discovery positions in plain English: why certain custodians matter, why others do not, and how proposed limits reflect proportionality. This is especially useful when translating technical search methodology into language suitable for opposing counsel—or the court.

Used carefully, LLMs help sharpen the *explanation* of discovery positions, not the positions themselves. They are best employed to clarify rationale, not to invent it.

### Expert Report Deconstruction and Cross-Examination Planning

Beyond summarizing expert reports, LLMs can break opinions into assumptions, inputs, methods, and conclusions, exposing where an expert relies on disputed facts or unexplained leaps. This lends itself naturally to cross-examination planning and Daubert-style challenges.

### Privilege Log Assistance and Consistency Checks

Within well-defined constraints, AI can assist in drafting privilege log descriptions, normalizing entries, and identifying inconsistencies across large logs. It can also flag entries that appear under- or over-described.

This use should be framed carefully: AI assists with form and consistency, not privilege determinations. Final judgment remains squarely with counsel.

### Voir Dire and Jury Instruction Framing

LLMs can help generate neutral, open-ended voir dire questions designed to surface attitudes about damages, corporations, law enforcement, or experts without argument or indoctrination. They can also help reorganize jury instructions into juror-friendly thematic groupings for counsel's internal use.

### Settlement and Mediation Preparation

AI tools can assist in preparing internal case assessments, summarizing strengths and weaknesses, identifying factual disputes, and drafting concise mediation narratives. They can also help battle-test positions by articulating how an opponent might frame the same facts.

This use is most effective when the model is instructed to argue *both sides*—a task lawyers often avoid but benefit from confronting early.

## Ten Examples with Prompts

An LLM can be a game-changer in litigation and trial preparation, going well beyond simple document drafting. Having discussed what these tools can do at a high level, the examples below

illustrate practical, effective ways to deploy them. These are not "magic words." Treat them as starting points and adapt the prompts to the facts, forum, and posture of your case.

## 1. Drafting Deposition Questions:

- **Prompt Example:** "You are a Texas trial lawyer. Using the facts I provide, generate a topic-organized deposition outline for a corporate designee in a products liability case involving an allegedly defective medical device. Focus on design specifications, quality control, complaint handling, and failure reports. Do not assume facts not supplied. Flag areas requiring document support."

- **Application:** Build a disciplined outline tied to the pleaded liability theories and the areas of inquiry fairly encompassed by the witness's role, designation, and knowledge under the Texas Rules of Civil Procedure. In Harvey, constrain the task to documents in the matter vault to keep the outline grounded in the record.

## 2. Discovery and ESI Protocols:

- **Prompt Example:** "Draft a meet-and-confer letter proposing a proportional ESI protocol for a federal employment discrimination case in Texas. Address data sources, custodians, date ranges, search methodology (with examples), metadata fields, production format, and privilege clawback. Explain the rationale for each choice in plain English."

  Tip: you might try uploading my free primer on ESI Protocols along with your prompt: http://www.craigball.com/ESIProtocol.pdf

- **Application:**
  Accelerate preparation of defensible ESI proposals and meet-and-confer correspondence. This is particularly effective in Harvey, where custodians and systems may already be defined within the matter.

## 3. Trial Preparation and Cross-Examination:

- **Prompt Example:** "Using only the materials I provide, prepare a cross-examination outline for a medical expert offering causation opinions in a personal injury case. Organize questions by assumptions, methodology, and conclusions. Identify potential bias or financial incentives. Cite to specific report pages or testimony where possible."

- **Application:** Develop targeted cross-examination outlines and issue-spot weaknesses without drifting beyond the record.

## 4. Case Summaries for Motion Work:

- **Prompt Example:** "Using only the cases I provide, summarize the holdings relevant to preservation and spoliation standards. Identify the legal rule, the court's reasoning, and

limits on the holding. Do not supply citations beyond the provided materials. End with a verification list."

- **Application:** Generate concise, record-bound summaries for motion drafting while constraining hallucinated authority.

### 5. Privilege Logs and Document Review:

- **Prompt Example:** "Create a privilege log template suitable for federal court, including fields for document date, author, recipients, privilege asserted, and description. Then review the attached draft entries for consistency and over- or under-description. Do not determine privilege; flag issues for attorney review."
- **Application:** Assist with formatting, consistency, and quality control in large privilege logs—never with privilege determinations themselves.

### 6. Voir Dire Preparation:

- **Prompt Example:** "Prepare a voir dire outline for a wrongful death case involving a commercial truck collision in Texas. Draft neutral, conversational questions designed to surface attitudes toward commercial drivers, corporate responsibility, and damages. Include follow-up questions and a brief note distinguishing cause challenges from strikes."
- **Application:** Develop voir dire questions that probe bias without argument or indoctrination.

### 7. Opening and Closing Argument Assistance:

- **Prompt Example:** "Based on the facts I provide, draft a preliminary opening statement outline for a breach of contract case involving delayed delivery of construction materials. Focus on chronology, responsibility, and damages. Write in plain language suitable for a jury. This is a draft for refinement, not a final script."
- **Application:** Generate themes, structure, and transitions—not finished courtroom rhetoric.

### 8. Crafting Jury Instructions:

- **Prompt Example:** "Using Texas Pattern Jury Charge concepts, organize the elements of a negligence claim—duty, breach, causation, and damages—into plain-English explanations for internal use. Do not draft instructions for filing. Identify areas requiring verification against the current TPJC."
- **Application:** Help counsel understand and organize jury issues before drafting or requesting formal instructions.

### 9. Case Chronologies:

- **Prompt Example:** "Using only the attached documents, build a dated chronology of key events relevant to notice and response. Cite each entry to a document and page. If dates or facts are uncertain, flag them."
- **Application:** Create reliable chronologies for motion practice, mediation, or trial preparation, built incrementally and verified.

### 10. Building Better Prompts:

- **Prompt Example:** "Given the following task and constraints, help me draft a prompt that will extract information from uploaded interrogatories and produce a table identifying, for each interrogatory, likely information sources, custodians, and appropriate objections under the Texas Rules of Civil Procedure. Include self-check instructions and a verification list."

- **Application:** Use the LLM to design disciplined prompts for complex tasks. LLMs are often remarkably effective at structuring their own instructions—provided you supervise the result.

## Ten Tips for Improved Prompts

The examples above show what LLMs can do in practice. The tips below focus on **how to ask** so the tool works the way a trial lawyer thinks. These are not rigid rules. You can get good results without following every one. But when outputs disappoint, the fixes are usually found here.

### 1. Precision Matters: Frame with Factual and Legal Specificity

Avoid generalities. Prompts grounded in concrete facts, posture, and governing law yield more useful results.

*Example:*
"Draft voir dire questions tailored for a Texas medical malpractice case involving an alleged delay in cancer diagnosis, where the plaintiff contends a breach of the applicable standard of care."

**Specificity clarifies intent and reduces irrelevant or generic output.**

### 2. Direct the LLM to Perform Defined Legal Tasks

Frame prompts around what you would ask a junior lawyer to do, not abstract research goals.

*Example:*
"Prepare a checklist for an FRCP Rule 26(f) conference in a complex products liability case, emphasizing ESI preservation, proportionality, and phased discovery."

**Task-focused prompts align outputs with litigation workflow rather than academic exposition. In Harvey, this is most effective when tied to the documents and custodians already present in the matter vault.**

## 3. Break Complex Requests Down into Logical Steps

Multifaceted prompts often fail because the model tries to do too much at once.

*Example:*
"First, list the elements of defamation under Texas law. Next, identify which elements are most commonly challenged at the pleading stage. Finally, outline arguments for dismissal based on those elements."

**Stepwise instructions reduce oversimplification and make it easier to spot errors.**

## 4. Instruct the Model to Assume a Role—With Limits

Assigning a role can improve focus and tone, but it should reflect how lawyers actually think, not fictional authority.

*Example:*
"You are a Texas trial lawyer evaluating competing positions on the scope of ESI production in a commercial contract dispute. Draft a neutral analysis suitable for a meet-and-confer."

**Role prompts improve relevance without implying that the model replaces judgment or authority.**

## 5. Request Multiple Variations for Strategic Flexibility

When optionality matters, ask for alternatives.

*Example:*
"Provide three alternative opening statement outlines for a wrongful death case: one emphasizing responsibility, one focusing on chronology, and one centered on damages."

**Variations expand strategic choices and often reveal framing you might not have considered.**

## 6. Use Iterative Refinement—But Restate What Matters

Refinement works best when you re-anchor the model to what matters most.

*Example:*
"Revise the following deposition outline to emphasize contradictions between the witness's testimony and prior internal emails. Maintain a neutral tone and cite source documents."

**Iteration sharpens outputs, but remember that earlier context may fall out of scope as sessions grow.**

### 7. Anchor Responses to Legal Frameworks—Not Speculation

Encourage analysis grounded in recognized legal standards rather than free-floating commentary.

*Example:*
"Summarize the elements of promissory estoppel under Texas law and identify common evidentiary proof offered for each element. Flag areas requiring independent verification."

**This reduces conjecture and keeps the analysis tethered to doctrine.**

### 8. Clarify the Desired Level of Detail and Tone

Tell the model who the audience is and how polished the output should be.

*Example:*
"Explain comparative responsibility in Texas premises-liability cases for a CLE audience—accurate, formal, but accessible to non-specialists."

**Clear tone instructions prevent over-lawyering or oversimplification.**

### 9. Use Conditional ("If–Then") Prompts for Nuanced Analysis

Conditional framing encourages balanced reasoning rather than one-sided conclusions.

*Example:*
"If a party fails to timely designate an expert under the Texas Rules of Civil Procedure, explain the risks of exclusion and the circumstances under which a court may allow late designation."

**This approach surfaces contingencies and exceptions that matter in practice.**

### 10. Require Verification and Source Limits

Do not assume the model will police itself.

*Example:*
"Explain spoliation standards applicable in Texas courts. Identify the governing principles and flag where independent review of current case law is required. Do not fabricate citations."

**Explicit verification instructions reduce hallucinations and reinforce your supervisory role. In Harvey, pair this with instructions to rely only on materials within the matter vault.**

## Understanding Limitations and Ethical Guardrails

While powerful, LLMs are not substitutes for legal expertise or professional judgment. Neither ChatGPT nor Harvey provides authoritative legal research, and neither should be treated as a source of verified law or fact. These tools do not reliably access live court dockets or proprietary research services such as Lexis, Westlaw, or PACER, and—as discussed—they can generate incorrect, incomplete, or fabricated information.

Of the more than 700 *reported* cases to date where generative AI produced hallucinated content (typically fake citations), there is a startling commonality to all: *the lawyer simply failed to verify the authorities cited.* The common thread is—let's not mince words—lazy, careless lawyering.

Accordingly, every legal citation, quotation, and factual assertion must be read, verified, and owned by counsel before use. Courts have made clear they will not accept "the AI did it" as an excuse. Think of these tools as junior lawyers: fast, tireless, and often insightful—but requiring supervision, skepticism, and correction.

Ethical use also requires attention to confidentiality and privilege. In ChatGPT, that means understanding and configuring privacy controls, avoiding public share links, and being deliberate about what information is uploaded. In Harvey, governance is stronger by design—matter-centric workspaces, access controls, retention policies, and audit trails—but those safeguards do not relieve lawyers of their independent duties of competence, supervision, and confidentiality under Texas ethics rules. Technology can assist compliance; it does not outsource responsibility.

Finally, remember that LLMs reason statistically, not legally. They recognize patterns in language, not truth. When the record is thin, ambiguous, or contested, the risk of confident but wrong output increases. The antidote is disciplined prompting, scoped inputs, and explicit verification requirements—habits already familiar to good trial lawyers.

## Final Thoughts for Trial Lawyers

As a law professor, longtime trial lawyer, and special master, I've found AI tools genuinely useful for early-stage drafting, ideation, organization, and breaking through inertia. I also concede the obvious: AI is no substitute for a skilled lawyer. It will not replace judgment, credibility, or the art of advocacy.

What it can do—when properly constrained—is accelerate routine work, surface patterns you might otherwise miss, and help you explore arguments more efficiently. Whether drafting deposition outlines, structuring chronologies, summarizing testimony, or preparing ESI protocols, LLMs can act as force multipliers that amplify what good lawyers already do well.

The lawyers who benefit most are not those who chase novelty, but those who treat AI the way they treat every other tool in practice: with curiosity, discipline, and skepticism. Embrace AI for efficiency and organization, and you may find it becomes an indispensable part of your litigation toolkit. Just remember: *every tool needs a craftsman*, and no amount of technology can replace the judgment and insight of a seasoned trial lawyer.

# Appendix: AI Prompt to Critique and Improve Keyword Searches

Adapted from Ball in Your Court, August 19, 2024

For more than two decades, I've wished for a tool that could critique keyword searches the way an experienced e-discovery professional would—flagging stop words, poor logic, overbreadth, and avoidable noise, and explaining *why* a query is likely to fail. Despite years of effort, nothing like it ever materialized.

Large language models finally make that idea practical.

Used properly, an LLM can function as a search-strategy reviewer—not a replacement for human judgment, but a disciplined assistant that explains why certain keywords or Boolean queries are inefficient, risky, or poorly suited to the platform in use. For lawyers who rely on lexical search (and most still do), this can be a meaningful upgrade.

**How to Use the Prompt**

1. Open ChatGPT (or an equivalent LLM).
2. Paste the prompt below into the message field **without submitting it yet**.
3. Use the **file-attachment ("+") control** next to the prompt field to upload a discrete list of keywords and Boolean queries for assessment (plain text, Word, or PDF).
4. Submit the prompt. If the response truncates, instruct the model to continue.

For best results, specify the **review platform** (e.g., Relativity/dtSearch, DISCO, Everlaw) and upload any relevant reference material (search syntax guides, stop word lists, ESI protocols, or primers).[2]

**Prompt: AI Review of Keyword and Boolean Search Strategy**

**Role:**
You are an e-discovery search consultant assisting litigation counsel.

**Objective:**
Evaluate the effectiveness of the attached keyword list and Boolean queries for use in litigation discovery. The goal is to improve recall and precision while avoiding unnecessary noise, overbreadth, and platform-specific pitfalls.

**Scope and Constraints:**
• Analyze only the uploaded keywords and queries.
• Do not assume facts not provided.

[2] I saw improved results when I uploaded pertinent additional information, such as my own Primer on Processing in E-Discovery found here.

• Do not fabricate platform rules—flag uncertainty where platform behavior varies.
• Focus on defensibility and practical improvement.

**Analysis Framework (apply to each term or query):**

1. **Stop-Word and Noise Risk**
   Identify terms likely to be ignored, diluted, or overly common, noting platform sensitivity where relevant.
2. **Synonyms, Variants, and Misspellings**
   Identify obvious linguistic variants, abbreviations, and common misspellings that may affect recall.
3. **Industry or Case-Specific Jargon**
   Flag missing domain-specific terminology suggested by the context of the queries.
4. **Boolean Logic and Structure**
   Assess operator usage, grouping, and proximity logic for internal consistency and likely effectiveness.
5. **Syntax and Platform Compatibility**
   Identify syntax issues that may fail or behave differently across common e-discovery platforms.
6. **Overbreadth and False-Hit Risk**
   Flag terms or constructions likely to generate disproportionate noise.
7. **Refinement Suggestions**
   Propose specific, defensible improvements to each query.

**Output Requirements:**
• Present results in a table with one row per keyword or query and one column per analysis category above.
• After the table, provide a short narrative summarizing common issues and best-practice guidance for refining lexical searches.
• End with a **Verification Checklist** identifying assumptions that require human confirmation.

**Tone:**
Analytical, conservative, and suitable for use in a defensible discovery workflow.

**Practice Notes**

This prompt is not intended to “optimize” searches automatically or replace testing against real data. Its value lies in explaining why searches fail, surfacing avoidable mistakes, and improving lawyers’ understanding of lexical search mechanics.

As with all AI-assisted work, outputs must be reviewed, tested, and owned by counsel. Keyword search remains a blunt instrument—but until something better is universally affordable, reliable, and repeatable, we should at least sharpen it.