Matt Whibley
THE VARTAZARIAN LAW FIRM

The Big 8

CONCEPTS EVERY LAWYER MUST UNDERSTAND TO USE AI COMPETENTLY AND ETHICALLY

There are eight core concepts every lawyer must understand to be efficient and competent with artificial intelligence: confirmation bias, token limits, hallucinations, context bleed, personalization, mode selection, prompt verification, and confidentiality. These concepts are non-negotiable. When applied consistently, AI becomes a practical drafting and analysis tool. When ignored, AI becomes unpredictable and unreliable.

A useful mental model is to treat AI as the smartest and most reckless person you will ever meet. It can produce high-quality work quickly. It will also agree with you when you are wrong, answer questions it cannot actually support, and mix contexts unless boundaries are imposed. The lawyer's role is to impose structure, verification, and discipline.

1. Confirmation bias and agreeableness

Most general-purpose AI systems are designed to be helpful and agreeable. In legal work, that design creates a predictable risk: The model tends to reinforce the theory it is given. When presented with a case theme or liability theory, AI will often supply supporting arguments while minimizing or omitting weaknesses.

This is not a subtle issue. If you ask AI whether your argument is strong, it will usually confirm that it is – unless instructed otherwise. For lawyers this behavior can be misleading and dangerous.

The most effective way to reduce this risk is direct instruction. Tell the model not to please you. A simple instruction – *"Answer objectively and neutrally. Do not attempt to agree with me. Flag uncertainty."* – materially improves output quality and reduces the likelihood of polished but misleading drafts. This instruction should be used consistently.

Agreeableness should be treated as the model's default setting and countered intentionally:

- Ask for the best defense case. "Assume you represent the defendant. Identify the strongest arguments against my position and the authorities supporting them."
- Demand a judicial lens. "Assume the judge is skeptical. Identify the most likely grounds for denial and what additional facts or authority would change the outcome."

AI should be used to expose weaknesses early, not to validate assumptions. When prompted correctly, it becomes a useful adversarial tool rather than a confirmation engine.

2. Token limits and truncation

Token limits are not a minor technical detail. They are a primary reason AI output fails silently in legal work. A "token" is approximately three-quarters of a word. Every AI model has a maximum context window – the total number of tokens it can consider in a single request. That window includes system instructions, chat history, prompts, uploaded documents, and the model's response. Most systems also impose a separate maximum output limit.

When the material provided exceeds the model's context window, the model does not reliably warn the user. Instead, it typically reads only a portion of what was uploaded, ignores the remainder, and guesses at what it did not process. The output may appear complete – organized headings, citations, confident tone – while critical facts, testimony, or entire sections of the record were never reviewed.

This failure pattern is particularly dangerous because it is not obvious. Lawyers often assume that if a document was uploaded, it was read. That assumption is frequently wrong.

Token limits apply not only to large files but also to long conversations. As a chat grows, earlier facts and instructions may fall outside the usable context. This is one reason AI responses can degrade mid-project or drift away from earlier assumptions without explanation.

Platform	Flagship model	Maximum context	Maximum output
OpenAI / ChatGPT	GPT-5 (Pro / API)	~400,000 tokens	~128,000 tokens
Google / Gemini	Gemini 3 Pro	~1,048,576 tokens	~65,536 tokens
Anthropic / Claude	Claude Sonnet / Opus 4.5	~200,000 tokens (up to 1M in beta)	~64,000 tokens
xAI / Grok	Grok 4 / Grok Heavy	~256,000 tokens	Provider-dependent

Current high-end token limits (December 2025)
(Model limits published by vendors; application interfaces may impose lower caps.)

Most token-limit failures can be avoided with a structured workflow:

- Chunk the record. Break large records into defined ranges (e.g., "Medical records 1–200," "201–400," "Deposition 1–75").
- Summarize before analysis. Ask for an accurate summary first. Begin analysis in a new chat using the summary as the working record.

• Reset context intentionally. Periodically ask: "Summarize everything established so far in a clean bullet list with defined terms." Paste that summary into a new chat and continue.
• Ask what was reviewed. "List the documents and page ranges you reviewed and identify any materials not reviewed." Vague answers indicate truncation.

We should assume that AI does not automatically read everything unless the workflow is designed to make that possible.

3. Hallucinations and citation discipline

AI systems sometimes produce incorrect facts, incorrect quotations, and nonexistent legal authority. In litigation, the highest-risk version of this problem is fabricated authority – a plausible-sounding case name, a citation that does not exist, or a quotation that is not in the cited case.

This behavior is driven largely by the model's tendency to satisfy the user's request. When asked for authority supporting a position, AI may generate something that looks correct rather than admit uncertainty.

The baseline rule is simple and mandatory: Never file or transmit AI-generated legal authority unless you have personally verified the case, the holding, and any quoted language.

Additional practices materially reduce risk:
• Restrict the universe. "Answer only using published California appellate decisions."
• Separate sourcing from analysis. "List the controlling statutes and cases with citations. Do not analyze yet."
• Require pinpoint citations. "Provide pinpoint citations for every proposition. If uncertain, say so."
• Limit quotations. Summaries are safer than quotations; quotations must be verified against the source.
• Cross-check internally. Paste the output into a fresh chat or different model and ask it to verify each citation and quoted passage for accuracy and proper citation format.

AI should be used to identify research paths and analytical structure. Final authority always comes from verified sources.

4. Context bleed and boundaries

"Context bleed" occurs when an AI model carries facts, assumptions, or instructions from one task into another where they no longer apply. It can appear as minor inaccuracies – an incorrect date or party – or as significant errors, such as importing facts from a different case, blending medical histories, or mixing arguments across motions.

AI does not know that you have switched cases unless you explicitly tell it. It also does not reliably recognize that two documents involve different parties unless they are clearly labeled. The model treats the entire conversation as a single working context unless instructed otherwise.

AI must therefore be treated as literal.
• If you switch cases, state that you are switching cases.
• If prior assumptions no longer apply, instruct the model to discard them.
• If a fact is revised, restate it and identify what it replaces.

Practical guardrails reduce context bleed significantly:
• Use one chat per matter. Do not mix cases in a single thread.
• Name documents uniquely (e.g., "Smith depo 10-12-25," "Kaiser records 1–250").
• Use defined terms for parties and dates ("Plaintiff," "Defendant," "Incident Date").
• Ask for a fact table before drafting: "Create a table of assumed facts; ask me to confirm or correct each."
• Start a new chat when the task changes (for example, moving from medical chronology to summary judgment argument).

5. Personalization

Personalization can materially improve AI reliability when used conservatively. The most effective personalization consists of default behavioral instructions that apply to every legal task.

Examples include:
• Preserve citations exactly as provided.
• Do not invent authority.
• Flag uncertainty rather than guessing.
• Ask clarifying questions when material facts are missing.
• Use a neutral, professional tone suitable for court filings.

These instructions do not replace case-specific prompts. They standardize behavior and reduce recurring errors.

Personalizing ChatGPT

- Open ChatGPT.
- Click the profile icon.
- Select Customize ChatGPT.
- In "How would you like ChatGPT to respond?" enter concise instructions such as:

"For legal tasks, respond objectively and neutrally. Do not speculate or attempt to agree with me. Preserve citations exactly. Do not invent authority. Flag uncertainty. Ask clarifying questions when material facts are missing. Use a professional tone appropriate for court filings."

Avoid placing confidential information in persistent instructions.

Personalizing Gemini

- Open Gemini.
- Open Settings.
- Locate Preferences or Custom instructions.
- Enter similar behavioral rules.
- Select a high-reasoning model (e.g., Gemini 3 Pro) before beginning legal analysis.

Gemini's output quality depends heavily on model selection at the outset. Lawyers should select the reasoning-focused model before uploading records or asking legal questions.

Personalizing Claude

- Open Claude
- Click on you profile
- Answer the question: "What personal preferences should Claude consider in responses?" with similar behavioral rules.

6. Model selection

Why model choice matters: Instant vs. high-reasoning modes

Not all AI models perform the same kind of work, even when given the same prompt. The difference is not just speed; it is how the model allocates effort. Understanding this distinction is critical for legal research and analysis.

In platforms like ChatGPT, "Instant" modes are optimized for responsiveness. High-reasoning modes (such as GPT-5 Pro, Gemini 3 Pro, or Claude Opus 4.5) are optimized for deliberation. Each has an appropriate use, and mixing them within the same workflow often produces unreliable results.

When you ask a legal research question in Instant mode, the model behaves like a fast drafting assistant. It commits quickly to a reasonable interpretation of the question and produces something that looks complete: a list, a short explanation, or a set of takeaways. It prioritizes plausibility and usefulness over careful framing. Nuance is compressed, assumptions are filled in smoothly, and the output often reads well – but it may gloss over ambiguity or unresolved issues that matter in real litigation.

High-reasoning modes behave differently. Given the same question, Pro mode spends more effort defining what must be resolved before answering. It decomposes the task into sub-questions, clarifies what the question actually turns on, and structures the response so it can be verified. Rather than jumping straight to an answer, it does more "pre-answer work": framing the issue, identifying decision points, and organizing the analysis in a way that mirrors how lawyers actually reason through problems.

This difference matters. For example, asking about the "scope" of an evidentiary rule in Instant mode will usually produce a quick, blended response. In Pro mode, the model is more likely to separate what the rule excludes, what it permits, and what factors courts treat as dispositive – making the output easier to audit and harder to misuse.

The practical rule is simple: Use Instant modes for brainstorming, rewriting, and first-pass drafting. Use high-reasoning modes for legal research, issue-spotting, and structured analysis. Generate ideas quickly, then move the work product into a reasoning-focused model for evaluation and refinement.

7. Prompt verification

Prompt verification is a simple technique for ensuring that you and the AI are working from the same understanding before any substantive work begins. It is most useful when a task is complex, the inputs are nuanced, or the outcome matters.

AI does not pause to confirm intent. If a prompt is even slightly ambiguous, the model will choose an interpretation and proceed confidently. That interpretation may not be the one you intended. Prompt verification prevents this by forcing alignment first.

A prompt that works consistently is:

"Do not work on this yet. Ask me three clarifying questions to make sure we are on the same page."

This instruction changes the interaction. Instead of guessing what you want, the model must identify what is unclear and resolve it before moving forward. The questions it asks often reveal ambiguities in scope, assumptions about the record, or differences in how the task is being understood.

In practice, prompt verification slows the process briefly at the beginning and saves time later. It reduces rewrites, prevents work based on incorrect assumptions, and produces output that more closely matches what you actually need.

Prompt verification mirrors good professional practice. Lawyers routinely clarify assignments before beginning work. Requiring AI to do the same turns it from a reactive text

generator into a more controlled and reliable tool.

The rule is straightforward: when the task matters, do not let the AI start working until you are certain you and the model are aligned.

8. Confidentiality, privilege, and work product

Federal courts have recognized that AI prompts and outputs created by counsel can constitute attorney work product.

In *Tremblay v. OpenAI, Inc.* (N.D. Cal. Aug. 8, 2024) 2024 WL 3748003, the court held that prompts crafted by counsel reflected attorney mental impressions and qualified as opinion work product. The court rejected a broad subject-matter waiver theory that would have treated prompts and outputs as fact work product. (*Id.* at *2–*3.)

In *Concord Music Group, Inc. v. Anthropic PBC* (N.D. Cal. May 23, 2025) 2025 WL 1482734, the court similarly treated undisclosed AI-generated materials created during a pre-suit investigation as protected work product and denied a motion to compel their production where plaintiffs represented that they had produced what they relied upon. (*Id.* at *1–*2.)

Most major AI platforms now offer private or non-retained chat modes. When enabled, conversations are not used for training and are retained only briefly for operational purposes – often approximately 30 days – before automatic deletion. If the user also deletes the conversation from their account interface, no persistent copy remains accessible.

From a litigation perspective, this reflects intentional data minimization. Once the provider's retention period expires, there is no continuing server-side record to produce. While lawyers should avoid categorical claims of immunity from discovery, using private chats, disabling history, and deleting conversations materially reduces the likelihood that AI interactions exist as discoverable materials at all.

Matt Whibley attended the University of Ottawa in Ontario, Canada, then went on to tour the world with a platinum-selling and Grammy-nominated Canadian punk rock band. He then attended Southwestern Law School in Los Angeles, earning over a dozen academic awards and scholarships and graduating Summa Cum Laude. He attended the Trial Lawyers College in 2017.