

The Fallacy of AI Confidentiality and Practical Safeguards

The rapid integration of generative artificial intelligence into pre-litigation research and case strategy has created a dangerous illusion that queries including documents uploaded into an AI model are as protected as a confidential consultation with a licensed attorney. In reality, uploading factual claims, litigation theories, or client secrets into commercial AI tools routinely exposes that information to disclosure and waives core evidentiary protections.

This chapter examines why information uploaded to AI platforms fails to meet the legal standards for privilege and outlines concrete, enforceable protocols to safeguard sensitive legal work product.

Evidentiary privileges specifically the attorney-client privilege and the work product doctrine are strictly construed exceptions to the general rule that the public is entitled to every person's evidence as illustrated in *re Grand Jury Subpoenas*, 318 F.3d 379, 384 [2d Cir. 2003]]. In *re Grand Jury Subpoenas*¹, the government argued that the work product doctrine applies only to materials prepared by or at the behest of counsel in anticipation of litigation. When a party inputs case facts or legal theories into a commercial generative AI model, the interaction fails every element required to establish privilege. As established in *United States v. Mejia*², the attorney-client privilege requires a communication:

- a. Between a client and an attorney;
- b. Intended to be, and kept, confidential; and
- c. Made for the primary purpose of obtaining or providing legal advice.

Commercial AI platforms fail on all three grounds:

- a. **No Attorney-Client relationship:** An AI platform holds no law license, owes no duties of loyalty or professional care, and is not a member of any bar. As reasoned in *In re OpenAI, Inc., Copyright Infringement Litig.*, discussing legal matters between non-attorneys creates no privilege³.

¹ *In re Grand Jury Subpoenas*, 318 F.3d 379, 384 [2d Cir. 2003]]

² *United States v. Mejia*, 655 F.3d 126, 132 (2d Cir. 2011)

³ *In re OpenAI, Inc., Copyright Infringement Litig.*, 802 F. Supp. 3d 688, 699 (S.D.N.Y. 2025)

- b. **Express disclaimer of legal advice:** Public commercial generative AI systems feature terms of service and underlying guardrails explicitly stating that the system cannot provide legal advice and recommending consultation with a qualified attorney. An individual cannot claim to use a tool to obtain legal advice when the tool disclaims that function.
- c. **Absence of confidentiality:** Prompts submitted to public platforms are subject to platform Privacy Policies that permit data collection, model training, and disclosure to third parties or regulatory authorities. Voluntary submission to a commercial third party destroys the expectation of confidentiality⁴.

The work product doctrine protects materials prepared *by or at the behest of counsel* in anticipation of litigation⁵. Unsupervised, self-initiated AI research conducted by a litigant or layperson is equivalent to independent internet searches, and it does not reflect counsel's mental impressions, strategy, or directives. Preexisting, unprivileged AI searches do not also become privileged simply because the user later emails or delivers the search history to an attorney⁶.

When a client feeds confidential attorney advice into an AI prompt for example, asking a chatbot to challenge a strategy their attorney proposed, they risk waiving privilege over the underlying attorney-client communications. In *re von Bulow*, the U.S Court of Appeal for the Second Circuit held that extrajudicial disclosures of confidential attorney-client communications do not trigger a broader subject matter waiver of the privilege unless those disclosures are deliberately used to gain a tactical advantage in litigation⁷. The third-party disclosure breaks the chain of confidentiality, exposing both the prompt and the original legal advice to discovery.

⁴ *United States v. Finazzo*, 682 F. App's 6, 16 [2d Cir. 2017]).

⁵ *In re Grand Jury Subpoenas*, 318 F.3d 379, 384 [2d Cir. 2003])

⁶ *United States v. Buyer*, 2023 WL 1381970 [S.D.N.Y. Jan. 31, 2023]; *United States v. Correia*, 468 F. Supp. 3d 618, 622 [S.D.N.Y. 2020]).

⁷ *In re von Bulow*, 828 F.2d 94, 101 [2d Cir. 1987]

In the case of *United States v. Heppner*, Bradley Heppner, CEO of Beneficient, was indicted for securities fraud, wire fraud, and related charges involving a scheme that allegedly misappropriated over \$300 million. Prior to his arrest, Heppner independently ran approximately 31 prompts through Anthropic's public AI model, Claude. Heppner's prompts incorporated details of the government's investigation, including strategic legal insights he had learned from his counsel. Heppner generated these documents for the stated purpose of analyzing his exposure and discussing the outputs with his defense team. The defense counsel did not direct or supervise these searches. Federal agents seized electronic devices containing these 31 AI-generated documents during the execution of a search warrant. The defense counsel asserted privilege over the files, leading the U.S. Attorney's Office for the Southern District of New York to move for an order declaring that no privilege attached. The government argued, and the court agreed, that the documents were discoverable based on three primary grounds⁸:

- a. **No Attorney Interaction:** Under *United States v. Ackert*, communications must occur with a licensed attorney or their agent. The AI platform is a commercial technology, not a licensed attorney⁹.
- b. **Lack of Confidentiality:** Anthropic's Privacy Policy and Terms of Service specify that the company retains user inputs and outputs, uses them for model training, and reserves the right to disclose data to governmental authorities.
- c. **No Work Product Shield:** Under *In re Grand Jury Subpoenas*, work product protection requires attorney involvement¹⁰. Because Heppner conducted the searches on his own initiative without counsel's direction or supervision, no work product protection attached.

In order to integrate AI capabilities without waiving privilege, legal teams and corporate clients must adopt institutional safeguards which include;

⁸ *United States v. Heppner*, 25 Cr. 503 (JSR)

⁹ *United States v. Ackert*, 169 F.3d 136 (2d Cir. 1999)

¹⁰ *In re Grand Jury Subpoenas*, 318 F.3d 379 (2d Cir. 2003)

1. Enterprise infrastructure & contractual architecture:

- a. **Zero-data retention contracts:** Do not use public consumer tiers. Procure enterprise AI instances with terms of service explicitly forbidding vendor logging, human review, or use of customer data for model training.
- b. **Opt-Out of external disclosures:** There is cautious need to ensure that agreements mandate strict confidentiality and restrict third-party disclosures, asserting that data remains sole property of the subscriber.

2. Strict protocol for counsel direction:

- a. **The Kovel & work-product framework:** All AI usage related to a pending or anticipated dispute must be initiated, directed, and supervised by a lawyer. In *United States v. Kovel*, court found that what is vital to privilege is that the communication is made in confidence for the purpose of obtaining legal advice from the lawyer, and that the privilege applies to non-lawyer employees or agents who act as a translator to help lawyers comprehend specialized subject matter¹¹.
- b. **Documenting directives:** Before running an AI query, a lawyer should document the legal strategy behind the research in an internal file note.

3. Client guidance & operational rules:

- a. **Clear prohibition on public chatbots:** Issue clear written policies prohibiting employees, executives, and clients from inputting legal advice, unreleased financial data, or case strategy into non-enterprise AI systems.
- b. **Prompt formatting & labeling:** When authorized enterprise AI tools are used under a lawyer's direction and the prompts should be explicit.

When evaluating the discoverability of AI-generated inputs and outputs during pretrial proceedings, courts apply the following evidentiary framework.

- a. **Unsupervised layperson use = unprivileged:** Independent AI searches conducted by a party without attorney direction are not protected by attorney-client or work product privileges.

¹¹ *United States v. Kovel*, 296 F.2d 918 [2d Cir. 1961]).

- b. **Subsequent delivery does not privilege:** Sending non-privileged AI prompts or results to a lawyer after the fact does not retroactively shield them¹². Merely transmitting non-privileged documents to a lawyer is not a privileged communication.
- c. **Public AI disclosures waive confidentiality:** Inputting attorney communications into public-facing AI tools waives the underlying privilege by disclosing secrets to a third party
- d. **Counsel supervision is required for protection:** Work product claims require proof that AI tools were utilized under the direction and strategy of legal counsel.

In conclusion, the evidentiary landscape surrounding artificial intelligence is governed by traditional legal doctrines applied to new technological realities: machine interfaces cannot substitute for licensed legal counsel, nor can commercial servers serve as confidential repositories for unprivileged research. As demonstrated in *United States v. Heppner*, a litigant's independent, unsupervised interactions with commercial GenAI tools no matter how deeply focused on pending legal matters or past legal advice fall entirely outside the protections of the attorney-client privilege and the work product doctrine.

The act of inputting sensitive case facts into public AI platforms operates as a voluntary third-party disclosure that destroys the expectation of confidentiality and risks waiving privilege over underlying legal strategy. Furthermore, transmitting the resulting digital outputs to defense counsel after the fact cannot retroactively transform non-privileged AI research into protected legal communications or work product.

To navigate these evidentiary risks during trial preparation, legal practitioners must enforce rigid, institutional safeguards. Protecting client confidences in the age of generative AI requires abandoning consumer-grade tools in favor of enterprise systems governed by zero-data retention contracts and explicit non-disclosure terms. Crucially, any analytical or research use of AI must be initiated, directed, and documented by licensed counsel under a clearly established work product protocol. In the absence of counsel's direct oversight and

¹² *United States v Buyer*, 2023 WL 1381970 (S.D.N.Y. 2023)

strict contractual data boundaries, inputs to AI systems will remain fully discoverable leaving both client communications and litigation strategy exposed to adverse inspection in open court.