您的 AI 提示詞是否可披露？每家公司和律師事務所都應了解的近期案例

Ruizheng (Tony) Wang 与 Califf T. Cooper

中文編譯：左涵湄、程錫佩

您的 ChatGPT、Claude 和 Gemini 提示詞及回復是否在證據開示（Discovery）中免於披露？近期的判決爲這一問題提供了啟示，揭示了您的 AI 聊天記錄在何時會成爲證據開示的對象，以及在何時仍屬於受保護的禁區。本文探討了現行規則如何應用於前沿生成式 AI，以及每家公司、企業法律部門和律師事務所必須採取哪些措施來避免“AI 證據開示災難”。

要了解 AI 證據開示方面的最新規則，必須關注《紐約時報》（NYT）與 OpenAI 之間的版權侵權訴訟。NYT 指控 OpenAI 未經許可或補償，不當地使用數百萬篇 NYT 受版權保護的文章來訓練大語言模型（LLM），從而構成大規模版權侵權。此案的核心問題在於，OpenAI 的數據抓取是否屬於“合理使用”，以及所産生的 ChatGPT 模型是否不當地重現了 NYT 受版權保護的內容。

在此背景下，美國地區法院法官 Ona T. Wang 做出的兩項近期證據開示裁決，爲 AI 提示詞和日誌何時具有潛在可披露性提供了深度參考。

在第一項裁決中，OpenAI 試圖強制開示《紐約時報》的內部“ChatExplorer”日誌，包括 NYT 員工向使用 OpenAI 模型的內部工具發送的提示詞及其收到的輸出[1]。OpenAI 主張這些日誌與“合理使用”和“實質性非侵權用途”相關。法院不同意並駁回強制開示的動議，理由是被告未能證明這些 ChatExplorer 日誌與案件需求相關（relevant）或相稱（proportional）。法官 Wang 解釋説，合理使用調查的重點是 OpenAI 對 NYT 受版權保護的作品的使用，而非 NYT 對 OpenAI 工具的下游後續使用。法院還認定，從相關著作權意義上看，NYT 對 ChatExplorer 的使用並不能證明市場損害，因爲 NYT 不能成爲其自身的“競爭替代品”。即便這些記錄有微薄的邊際相關性，要從逾 8 萬條日誌中逐一審查哪些受保密特權保護，其負擔也使該請求明顯不具比例性。

法官 Wang 在同一訴訟中的後續裁決呈現了問題的另一面。並非是 OpenAI 試圖獲取 NYT 的內部 AI 使用日誌，而是 NYT 要求 OpenAI 提供其消費者 ChatGPT 輸出日誌，以證明 AI 確實重現了 NYT 受版權保護的文章[2]。在該裁決中，法官 Wang 駁回了 OpenAI 的複議動議（motion for reconsideration），強制該公司提供 2000 萬條去標識化的消費者 ChatGPT 日誌樣本。法院認定，這些日誌與案件的核心實質問題（輸出是否複製 NYT 的

[1] *In re OpenAI, Inc., Copy. Infringement Litig.*, 800 F. Supp. 3d 602, 606 (S.D.N.Y. September 19, 2025).
[2] *In re OpenAI, Inc., Copy. Infringement Litig.*, No. 25-MD-3143 (SHS) (OTW), 2025 WL 3468036, at *1 (S.D.N.Y. Dec. 2, 2025).

作品）高度相關並且相稱，因爲 2000 萬條日誌僅佔 OpenAI 保留的數百億條日誌的不到 0.05%。這些輸出也可能與 OpenAI 自身的抗辯（包括合理使用和實質性非侵權用途）相關。法院進一步認定，去標識化工作已基本完成，隱私保護措施也已到位。

在版權侵權背景之外，近期 *United States v. Heppner* 案的裁決爲律師尚未介入時的情況提供了啟示。Bradley Heppner 是一起涉及涉嫌證券和電信欺詐的聯邦刑事案件的被告 [3]。在被捕前，Heppner 曾使用 Anthropic 的 Claude 生成關於政府調查和可能辯護策略的文件。在該案件中，美國地區法院法官 Rakoff 裁定，這些生成的文件不受律師-客戶特權（Attorney-Client Privilege）或工作成果原則（Work-Product Doctrine）的保護。法院的理由是，Claude 不是律師，該通信不具有保密性，並且這些文件並非由律師準備或在其指導下準備。該意見還指出了 Anthropic 的條款和隱私披露，包括輸入和輸出可能被收集、用於訓練 Claude 以及被披露給第三方。法院確實指出，如果是律師指導客戶使用 Claude，可能會呈現一個更複雜的特權問題，但本案並非如此。該裁決表明，當律師-客戶特權和工作成果的論點不成立時，客戶在沒有律師參與的情況下對 AI 的使用很可能會被納入證據開示範圍

這些案例給了我們重要的啟示：AI 提示詞並非因爲存在就自動具有可披露性。當事人仍必須證明提示詞、輸出或日誌對實際的訴訟請求或抗辯至關重要，並且對其進行收集和審查的負擔是合理的。與案件實質過於疏遠或審查負擔過重的 AI 日誌，對其提出的寬泛請求可能會失敗。另一方面，當存在以下情況時，AI 提示詞更可能可披露：1）與訴訟請求或辯護相關；2）針對這些 AI 提示詞的請求與案件需求相稱；3）不受律師-客戶特權或工作成果原則保護；4）並非在律師指導下創建；或是 5）AI 工具的條款或設置破壞了保密性。歸根結底，傳統證據開示規則同樣適用於由尖端 LLM 和前沿 AI 平台生成的數據。

這些案例提供了一個極其清晰的警示，即職業責任嚴格延伸至律所及其客戶對 AI 的使用方式。提示詞和輸出屬於電子存儲信息（ESI），受到與電子郵件或短信相同的職業倫理和證據開示義務約束。美國律師協會（ABA）《職業行爲示範規則》第 1.1 條第 8 款注釋要求律師必須了解與相關技術關聯的益處和風險。如果律師將敏感客戶數據輸入到會利用用戶輸入進行訓練或與第三方共享信息的公共消費者 AI 工具中，則面臨違反倫理規則的風險。更糟糕的是，他們可能會完全喪失律師-客戶特權，這意味着這些提示詞將失去保護，並可被對方律師要求開示。

《示範規則》第 3.3 條和第 3.4 條也同樣適用，因爲律師不得向法庭提交虛假的 AI 生成材料，不得阻礙他人獲取證據，也不得允許相關材料丟失或被毀。如果 AI 提示詞或輸出與訴訟請求相關，律師必須在合理預見到會發生訴訟的情況下立即採取措施予以保存。

---

[3] *U.S. v. Heppner*, No. 25 CR. 503 (JSR), 2026 WL 436479, at *1 (S.D.N.Y. Feb. 17, 2026).

這意味着要主動建議客戶關閉 AI 聊天日誌的自動刪除功能（這個問題在 OpenAI 訴訟中出現過，當時法院不得不介入以阻止消費者輸出日誌的常規刪除）。正如最近的新聞報導所表明的那樣，律師不能向法庭提交虛假的 AI 生成材料（例如 AI“幻覺”捏造的判例法）。律師事務所必須在依賴 LLM 生成的任何輸出之前對其進行獨立驗證。

最後，ABA 的《示範規則》第 5.1 條和第 5.3 條要求律所和主管律師實施合理措施，以確保律師和非律師的助理以符合職業義務的方式使用 AI。“不知情”不能作爲抗辯理由。主管律師和律師事務所必須實施清晰的書面政策以及合理的措施，以確保律師和非律師的助理以保護客戶數據並遵守證據開示義務的方式使用 AI。

AI 提示詞既不自動具有可披露性，也不自動免於披露。當 AI 提示詞直接涉及案件實質、與案件相稱且不受特權保護時，它們就是可被披露的。企業此時就應該實施嚴格的 AI 保留和使用政策，以免日後被迫交出其聊天記錄。