您的 AI 提示词是否可披露？每家公司和律师事务所都应了解的近期案例

Ruizheng (Tony) Wang 与 Califf T. Cooper

中文编译：左涵湄、程锡佩

您的 ChatGPT、Claude 和 Gemini 提示词及回复是否在证据开示（Discovery）中免于披露？近期的判决为这一问题提供了启示，揭示了您的 AI 聊天记录在何时会成为证据的开示对象，以及在何时仍属于受保护的禁区。本文探讨了现行规则如何应用于前沿生成式 AI，以及每家公司、企业法律部门和律师事务所必须采取哪些措施来避免“AI 证据开示灾难”。

要了解 AI 证据开示方面的最新规则，必须关注《纽约时报》（NYT）与 OpenAI 之间的版权侵权诉讼。NYT 指控 OpenAI 未经许可或补偿，不当地使用数百万篇 NYT 受版权保护的文章来训练大语言模型（LLM），从而构成大规模版权侵权。此案的核心问题在于，OpenAI 的数据抓取是否属于“合理使用”，以及所产生的 ChatGPT 模型是否不当地重现了 NYT 受版权保护的内容。

在此背景下，美国地区法院法官 Ona T. Wang 做出的两项近期证据开示裁决，为 AI 提示词和日志何时具有潜在可披露性提供了深度参考。

在第一项裁决中，OpenAI 试图强制开示《纽约时报》的内部“ChatExplorer”日志，包括 NYT 员工向使用 OpenAI 模型的内部工具发送的提示词及其收到的输出[1]。OpenAI 主张这些日志与“合理使用”和“实质性非侵权用途”相关。法院不同意并驳回强制开示的动议，理由是被告未能证明这些 ChatExplorer 日志与案件需求相关（relevant）或相称（proportional）。法官 Wang 解释说，合理使用调查的重点是 OpenAI 对 NYT 受版权保护的作品的使用，而非 NYT 对 OpenAI 工具的下游后续使用。法院还认定，从相关著作权意义上看，NYT 对 ChatExplorer 的使用并不能证明市场损害，因为 NYT 不能成为其自身的“竞争替代品”。即便这些记录有微薄的边际相关性，要从逾 8 万条日志中逐一审查哪些受保密特权保护，其负担也使该请求明显不具比例性。

法官 Wang 在同一诉讼中的后续裁决呈现了问题的另一面。并非是 OpenAI 试图获取 NYT 的内部 AI 使用日志，而是 NYT 要求 OpenAI 提供其消费者 ChatGPT 输出日志，以证明 AI 确实重现了 NYT 受版权保护的文章[2]。在该裁决中，法官 Wang 驳回了 OpenAI 的复议动议（motion for reconsideration），强制该公司提供 2000 万条去标识化的消费者

---

[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).

ChatGPT 日志样本。法院认定，这些日志与案件的核心实质问题（输出是否复制 NYT 的作品）高度相关并且相称，因为 2000 万条日志仅占 OpenAI 保留的数百亿条日志的不到 0.05%。这些输出也可能与 OpenAI 自身的抗辩（包括合理使用和实质性非侵权用途）相关。法院进一步认定，去标识化工作已基本完成，隐私保护措施也已到位。

在版权侵权背景之外，近期 *United States v. Heppner* 案的裁决为律师尚未介入时的情况提供了启示。Bradley Heppner 是一起涉及涉嫌证券和电信欺诈的联邦刑事案件的被告。[3]在被捕前，Heppner 曾使用 Anthropic 的 Claude 生成关于政府调查和可能辩护策略的文件。在该案件中，美国地区法院法官 Rakoff 裁定，这些生成的文件不受律师-客户特权（Attorney-Client Privilege）或工作成果原则（Work-Product Doctrine）的保护。法院的理由是，Claude 不是律师，该通信不具有保密性，并且这些文件并非由律师准备或在其指导下准备。该意见还指出了 Anthropic 的条款和隐私披露，包括输入和输出可能被收集、用于训练 Claude 以及被披露给第三方。法院确实指出，如果是律师指导客户使用 Claude，可能会呈现一个更复杂的特权问题，但本案并非如此。该裁决表明，当律师-客户特权和工作成果的论点不成立时，客户在没有律师参与的情况下对 AI 的使用很可能会被纳入证据开示范围。

这些案例给了我们重要的启示：AI 提示词并非因为存在就自动具有可披露性。当事人仍必须证明提示词、输出或日志对实际的诉讼请求或抗辩至关重要，并且对其进行收集和审查的负担是合理的。与案件实质过于疏远或审查负担过重的 AI 日志，对其提出的宽泛请求可能会失败。另一方面，当存在以下情况时，AI 提示词更可能可披露：1）与诉讼请求或辩护相关；2）针对这些 AI 提示词的请求与案件需求相称；3）不受律师-客户特权或工作成果原则保护；4）并非在律师指导下创建；或是 5）AI 工具的条款或设置破坏了保密性。归根结底，传统证据开示规则同样适用于由尖端 LLM 和前沿 AI 平台生成的数据。

这些案例提供了一个极其清晰的警示，即职业责任严格延伸至律所及其客户对 AI 的使用方式。提示词和输出属于电子存储信息（ESI），受到与电子邮件或短信相同的职业伦理和证据开示义务约束。美国律师协会（ABA）《职业行为示范规则》第 1.1 条第 8 款注释要求律师必须了解与相关技术关联的益处和风险。如果律师将敏感客户数据输入到会利用用户输入进行训练或与第三方共享信息的公共消费者 AI 工具中，则面临违反伦理规则的风险。更糟糕的是，他们可能会完全丧失律师-客户特权，这意味着这些提示词将失去保护，并可被对方律师要求开示。

《示范规则》第 3.3 条和第 3.4 条也同样适用，因为律师不得向法庭提交虚假的 AI 生成材料，不得阻碍他人获取证据，也不得允许相关材料丢失或被毁。如果 AI 提示词或输

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

出与诉讼请求相关，律师必须在合理预见到会发生诉讼的情况下立即采取措施予以保存。这意味着要主动建议客户关闭 AI 聊天日志的自动删除功能（这个问题在 OpenAI 诉讼中出现过，当时法院不得不介入以阻止消费者输出日志的常规删除）。正如最近的新闻报道所表明的那样，律师不能向法庭提交虚假的 AI 生成材料（例如 AI“幻觉”捏造的判例法）。律师事务所必须在依赖 LLM 生成的任何输出之前对其进行独立验证。

最后，ABA 的《示范规则》第 5.1 条和第 5.3 条要求律所和主管律师实施合理措施，以确保律师和非律师的助理以符合职业义务的方式使用 AI。“不知情”不能作为抗辩理由。主管律师和律师事务所必须实施清晰的书面政策以及合理的措施，以确保律师和非律师的助理以保护客户数据并遵守证据开示义务的方式使用 AI。

AI 提示词既不自动具有可披露性，也不自动免于披露。当 AI 提示词直接涉及案件实质、与案件相称且不受特权保护时，它们就是可被披露的。企业此时就应该实施严格的 AI 保留和使用政策，以免日后被迫交出其聊天记录。