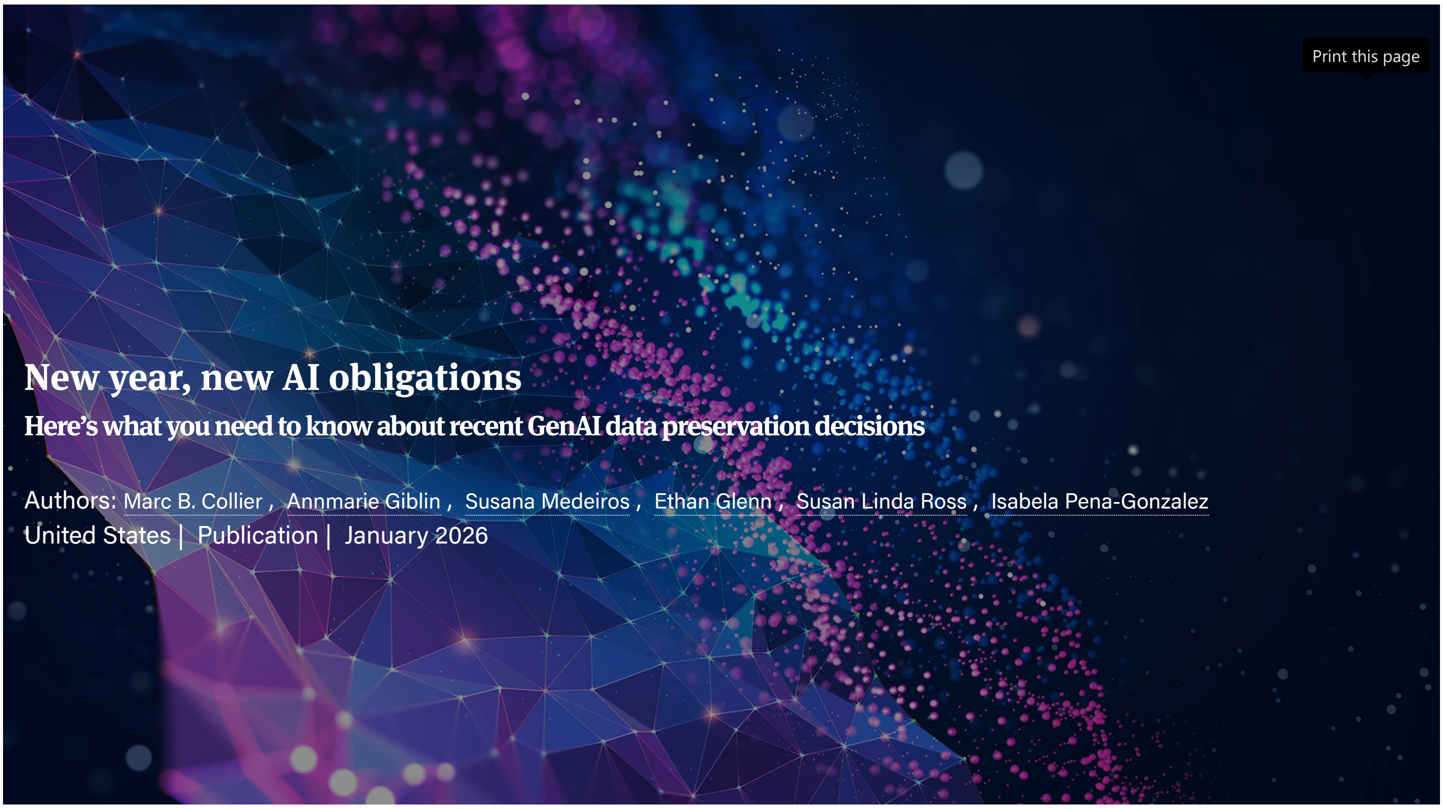
NORTON ROSE FULBRIGHT



New year, new AI obligations

Here’s what you need to know about recent GenAI data preservation decisions

Authors: Marc B. Collier , Annmarie Giblin , Susana Medeiros , Ethan Glenn , Susan Linda Ross , Isabela Pena-Gonzalez

United States | Publication | January 2026

The fast adoption of generative artificial intelligence (GenAI) has come with more than a few growing pains. Many industries have had to quickly adapt to this new technology at the same time that courts are issuing GenAI-related decisions in courtrooms across the country that could have broad-reaching effects on companies who use GenAI, even those companies that are not currently in litigation. This article summarizes some recent GenAI-related decisions and discusses their implications for companies.

As background, for many GenAI systems and in particular Large Language Model (LLM) based systems, when a user creates an input—a prompt—to the GenAI system, that GenAI creates a text and media response based on the data it was trained on, which is referred to as an output. Together, those prompts and outputs are generally stored in a log format. The treatment of those logs raises questions about whether, how and by whom they should be preserved for litigation.

These questions came to a head recently in *In re: OpenAI Inc. Copyright Infringement Litigation*, which is currently pending in the Southern District of New York. In that case, *The New York Times* (NYT) and other publications and authors sued OpenAI and another defendant, alleging that, among other things, OpenAI used copyrighted content to train OpenAI’s ChatGPT product without permission or payment. In the course of resolving several discovery disputes related to the production of GenAI outputs, the court has repeatedly given guidance for how companies operating nationwide should treat GenAI inputs and outputs going forward. Specifically, the court issued orders addressing and circumscribing preservation and production obligations for GenAI data in litigation.

This article discusses three of the court’s 2025 and 2026 decisions—the latest issued less than three weeks ago—that touch on one specific issue that will affect any company that uses GenAI: the reasonable preservation of GenAI inputs and outputs. This article provides a summary of the key points from the three orders and a detailed discussion of each. It also identifies key takeaways that legal practitioners should consider when managing GenAI data, and suggests best practices for how, in light of these orders, companies who employ GenAI systems can start their new year off on the right foot and ensure that they are reasonably preserving and managing that data.

Summary of key points from the court’s orders

- The court ordered OpenAI to preserve *and* segregate *all* output log data that would otherwise be deleted pursuant to OpenAI’s retention policies.
- The court ordered OpenAI to preserve ChatGPT output log data even where OpenAI argued that the data would be deleted to comply with consumers’ deletion requests under various privacy laws.

- The court denied OpenAI's Motion for Reconsideration and ordered OpenAI to produce a sample of 20 million retained, de-identified consumer ChatGPT output logs (the Logs) to the plaintiffs.
- The court held that the production of the 20 million logs was both relevant and proportional to the needs of the case.
- The court recognized the privacy considerations at issue in the case and determined that there were adequate layers of protection reasonably mitigating existing privacy concerns.
- The court determined that the production of the 20 million logs in de-identified form adequately balanced the privacy rights of users.

May 2025 order: Preserve and segregate all output log data

During discovery, NYT moved for an order requiring OpenAI to preserve ChatGPT user conversation logs, arguing this data could support its infringement claims. OpenAI opposed the motion, citing the burden of retaining billions of conversations and the impossibility of searching the conversations it did save. OpenAI also raised user privacy expectations and compliance with privacy laws requiring data deletion upon user request, arguing that ChatGPT users expected their chats to be unavailable after they were deleted. The company's own data retention policies promised users that if a user deleted a chat, the chat would be removed from the user's account immediately and scheduled for permanent deletion from OpenAI systems within 30 days (absent a legal or security reason to preserve it).

On May 13, 2025, US Magistrate Judge Ona T. Wang for the Southern District of New York granted NYT's motion and ordered OpenAI "to *preserve* and *segregate* all output log data that would otherwise be deleted (based on user preferences or deletion requests subject to privacy laws in the US and around the world) on a going forward basis until further order of the court."[1] This order effectively treats GenAI prompts and outputs as discoverable "documents" for purposes of litigation that are subject to preservation obligations. The order also clarifies that litigants are required to preserve not only AI outputs, but also the underlying datasets and algorithms. The court emphasized that traditional notions of relevance apply to preservation obligations regarding AI content, and that parties cannot rely on routine deletion policies, user privacy preferences or deletion requests governed by privacy laws to avoid their obligation to preserve potentially relevant information related to the parties' claims or defenses once litigation is reasonably anticipated. The court's order confirms that parties' preservation duties include GenAI inputs and outputs. Should other courts adopt a similar approach, organizations should be prepared to implement reasonable methods to identify and preserve potentially discoverable GenAI content.

The court also noted that the principles of proportionality similarly apply to GenAI, leaving the door open for litigants to consider whether to negotiate with the opposing party or file a motion with the court that the scale and scope of GenAI information it is preserving, much of which may be of limited potential relevance to the claims and defenses, is disproportionate.

December 2025 order: Produce 20 million de-identified output logs

On November 7, 2025, the court directed OpenAI to produce a sample of the tens of billions of consumer ChatGPT logs that OpenAI had retained as a result of the May 2025 order. That sample size was set at 20 million retained, de-identified consumer ChatGPT output logs.

On December 2, 2025, the court denied OpenAI's request to reconsider the Court's November production order, finding that the production of the 20 million ChatGPT Logs was "relevant and proportional" to the needs of the case.[2] In particular, the court stated that the output logs were relevant because they may contain instances where the plaintiffs' copyrighted works may have been reproduced in whole or in part (and thus be evidence for the relevant claims). Additionally, the court reasoned that even the logs where reproductions of the plaintiffs' works were not contained may still prove to be relevant to OpenAI's fair use defense—specifically, to its calculation of damages. In light of that, the court determined that the sample of logs were "clearly relevant" to the extent they contain "partial or whole reproductions of the copyrighted works," and were relevant to the defendant's affirmative defenses "to the extent that they contain other user activity."[3]

The court reasoned that production of 20 million ChatGPT logs—as opposed to another sample number—was proportional to the needs of the case because, in part, "[t]he total universe of retained consumer output logs is in the tens of billions. The 20 million sample here represents less than 0.05 percent of the total logs that OpenAI has retained in the ordinary course of business."[4] Given the amount of log retention and de-identification OpenAI had already conducted, the court believed the burden on OpenAI to produce those logs would also be minimal.

The court emphasized that, while consumer privacy was a factor that it considered in making its decision on whether to order production of the logs, it was "just one factor in assessing the proportionality of discovery."[5] Given the de-identification of the logs, the protective order in the case and the designation of the logs as "attorneys' eyes only," the court determined that consumers' privacy rights were adequately protected, mitigating any existing privacy concerns.

January 2026 order: Ordering production of 20 million de-identified Logs adequately balanced users privacy rights

After the court issued its December 2025 order discussed above, OpenAI objected to the ruling on the grounds that it inadequately balanced user's privacy considerations with the potential relevance of the documents. The court conducted additional analysis and found that user's privacy considerations were properly balanced in light of privacy protections that were already in place.[6]

Specifically, the court based its decision on the fact that the users at issue here had voluntarily disclosed their conversations, and those disclosed conversations were retained in OpenAI's normal course of business. Additionally, there was no allegation here that OpenAI had illegally obtained the conversations. Therefore, the court determined that the privacy rights of users were different than the users in the circumstances OpenAI cited, and production adequately balanced their privacy rights to those voluntarily-disclosed conversations.

OpenAI had also argued the court's December 2025 order was erroneous because it rejected OpenAI's request to run relevant search terms against the 20 million de-identified logs—and thus only produce a sub-set of those logs—instead of producing all 20 million logs at issue. The court rejected that argument, finding that there was no requirement that the court order the least-burdensome discovery possible and that it was not erroneous to order production of all 20 million logs without additional searches.

Conclusion: Practical takeaways and proposed best practices

- Review and, as needed, revise your company's data management plan, including any relevant updates to information governance, record retention and legal hold policies and training, as it relates to GenAI inputs and outputs, including:
  - Evaluating your company's preservation practices that account for the technology's complexity and its potential relevance to litigation. The failure to take reasonable steps to preserve potentially relevant GenAI data could result in spoliation sanctions.
  - Evaluating retention of GenAI inputs and outputs. If GenAI information is retained long-term (whether for record retention or business need or otherwise), it may be subject to preservation and discovery obligations.[7]
  - If your company routinely disposes of GenAI inputs and outputs, ensuring your company can take reasonable steps to suspend disposition of this data to comply with legal hold and preservation requirements.
  - Providing guidance to employees that information entered into GenAI systems may be subject to discovery.
  - In light of the order from the Southern District of New York, consider the intersection of preservation obligations with privacy compliance. As a general matter, and as illustrated here, a company's duty to preserve trumps normal disposition practices, including disposition of data past-retention, and compliance with data deletion requests and data minimization requirements under privacy laws.[8] Consider whether your GenAI system(s) are likely to contain sensitive information, including information subject to privacy laws. Ensure that you have a process for suspending any data deletion requests or routine deletion where a legal hold applies.
- Consider implementing or updating your AI Governance program.
  - Implement an AI Governance program—or if one already exists, review and revise it as needed—to conform with the court's guidance and new laws going into effect this year;[9] Organizations should use this case as an opportunity to train employees on GenAI usage guidelines, privacy expectations (or lack thereof) and limits that the organization might place on its use (for example, restrictions—such as prohibiting input of confidential client information into commercial GenAI products).
  - Train employees on new GenAI-related obligations.
  - Employees should understand that the information they input into GenAI may be subject to review by your company. Where permissible in their local jurisdiction, clarify to your employees that they have no expectation of privacy in the data they input into GenAI.

- Consider updated e-Discovery and litigation strategies for handling GenAI preservation and discovery requests, including:
  - Whether there is GenAI data that is likely to be potentially relevant across many cases and thus may frequently be subject to legal hold and discovery requests – this likelihood will influence the importance and value of any strategic decisions your company may make on how to handle GenAI data in litigation.
  - Whether your company uses proprietary in-house GenAI and/or third party licensed GenAI, which may influence the level of control or capabilities you may have from a preservation and collection perspective. Where third parties are retaining GenAI data on behalf of your company, this retention may include considering whether additional steps may be needed to preserve potentially relevant information held by third parties.
  - Whether GenAI data is organized by custodian or any retention settings that may influence how and how quickly your company needs to take steps to preserve.
  - Whether you may need to update or include hold notice language that captures GenAI data.
  - Whether your company has a specific strategy around negotiating preservation and disclosure of GenAI data (both internally and as part of outside counsel guidelines). For example, litigants may want to discuss with opponents early on limitations around the discovery of GenAI information, including proportionality, potential costs and burdens, a defined scope of relevant prompts and outputs and key custodians related to GenAI specifically.
  - Whether your company tracks or classifies GenAI data that may be subject to a data deletion request, contains sensitive personal information or relates to custodians outside the United States where local privacy laws may further complicate preservation, collection and disclosure. This classification will likely factor into your existing strategies for managing sensitive information in e-Discovery (for example, proportionality arguments, phased discovery, protective orders, redaction protocols and de-identification where feasible).
  - As with all new technology, there are unique preservation, collection, review and disclosure considerations for GenAI. Common questions such as who is the 'custodian' of the data, or what constitutes a document 'family' in the GenAI context where inputs and outputs relate to both chats, meeting transcripts and documents, can all further complicate potential litigation strategies.
  - Whether to update e-Discovery guidelines, checklists and outside counsel guidelines, in light of these considerations.

More and more companies are rapidly investing in, using and planning to grow their use of GenAI. Given the increased use of, and significant volume of information generated by, these tools, GenAI may present outsized litigation challenges.

Our team has significant experience dealing with GenAI information, from both a governance perspective on the proper management of GenAI from its creation to storage and disposition, and a litigation and e-Discovery perspective, including strategies for preservation, collection, negotiation and disclosure. For any questions regarding issues relating to those matters, please contact us.

Footnotes

[1] *In re: OpenAI, Inc., Copyright Infringement Litigation*

[2] *In re OpenAI, Inc., Copyright Infringement Litig.*, No. 25-MD-3143 (SHS) (OTW), 2025 WL 3468036, at *2 (S.D.N.Y. Dec. 2, 2025).

[3] *Id.* at *3

[4] *Id.*

[5] *Id.*

[6] *In re OpenAI, Inc., Copyright Infringement Litig.*, No. 25-MD-3143 (SHS) (OTW), 2026 WL 21676, at *2 (S.D.N.Y. Jan. 5, 2026)

[7] For example, California's employment regulation (Cal. Code Regs., title 2, Section 11008) has a 4-year retention requirement that includes all "automated-decision system data, and other records created or received by the employer or other covered entity dealing with any employment practice and affecting any employment benefit of any applicant or employee."

[8] For example, under the California Consumer Privacy Act (CCPA), covered businesses are not required to comply with a California consumer's request to delete their personal information where the information is "reasonably necessary [to]. . . [c]omply with a legal obligation."

[9] *See, e.g.*, "The Texas Responsible AI Governance Act: What your company needs to know before January 1"; *see also* "The New York Responsible AI Safety and Education (RAISE) Act: What you need to know."

Contacts

Marc B. Collier

Head of Litigation and Disputes, Austin

marc.collier@nortonrosefulbright.com

Austin

T: +1 512 536 4549

Annmarie Giblin

Partner

annmarie.giblin@nortonrosefulbright.com

New York

T: +1 212 318 3080

Susana Medeiros

Partner

susana.medeiros@nortonrosefulbright.com

New York

T: +1 212 318 3044

Ethan Glenn

Senior Counsel

ethan.glenn@nortonrosefulbright.com

Austin

T: +1 512 536 2437

Susan Linda Ross

Senior Counsel

susan.ross@nortonrosefulbright.com

New York

T: +1 212 318 3280

Isabela Pena-Gonzalez

Associate

isabela.pena-gonzalez@nortonrosefulbright.com

Houston

T: +1 713 651 3669

Practice areas:

Artificial intelligence (AI) | eDiscovery | Cybersecurity and data privacy | Intellectual property | Litigation and disputes | Technology transactions

Industry:

Technology