

The Files are in the Computer: Copyright, Memorization, and Generative AI

Anonymous Authors¹

Abstract

A central issue in copyright lawsuits against generative-AI companies is the degree to which a generative-AI model does or does not “memorize” the data it was trained on. Unfortunately, the debate has been clouded by ambiguity over what “memorization” is, leading to legal debates in which participants often talk past one another. We attempt to bring clarity to the conversation over memorization.

1. Introduction

The week between Christmas and New Year’s Eve is usually a slow news week, but not this year, the year that ChatGPT ate the world (Chat GPT Is Eating the World; OpenAI, 2022; 2023). On December 27, 2023, *The New York Times* filed a massive copyright-infringement lawsuit against Microsoft and OpenAI, alleging that Bing Copilot and ChatGPT constituted “massive copyright infringement” (New York Times Company v. Microsoft, p. 74) In particular, the *Times* alleged that these models had “memorized” large quantities of *Times* articles (Carlini et al., 2021; 2023a;b; Nasr et al., 2023; Zhang et al., 2020). When prompted with text from a *Times* article,¹ ChatGPT would output lengthy passages from the article, hundreds of words, varying only in a few scattered portions. (See Figure 1.)

To the *Times* and its lawyers, these examples of “memorization” were blatant copyright infringement. But to OpenAI and its defenders, there was nothing to see here. OpenAI responded, both in court and online, that these examples were “adversarial, not normal usage patterns.” On this view, the prompts the *Times* used were the actual cause of the resulting copying, not evidence that copying had happened at some point within the technology powering ChatGPT. As

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

¹The prompts ranged in length from a sentence to several paragraphs. See New York Times Company v. Microsoft Complaint at Exhibit J.

Output from GPT-4:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

“Nobody wanted to upset the industry,” said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees medallions. “Nobody wanted to kill the golden goose.”

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund key initiatives.

During that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required borrowers to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

When the market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Actual text from NYTimes:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

“Nobody wanted to upset the industry,” said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees cabs. “Nobody wanted to kill the golden goose.”

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund priorities.

Under Mr. Bloomberg and Mr. de Blasio, the city made more than \$855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

But during that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required them to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

Figure 1. Memorized *New York Times* article from ChatGPT

economist Tyler Cowen put it, in mocking the *Times*’s argument, one could equally well say that a toothpick infringes:

If you stare at just the exact right part of the toothpick, and measure the length from the tip, expressed in terms of the appropriate unit and converted into binary, and then translated into English, you can find any message you want. You just have to pinpoint your gaze very very exactly (I call this “a prompt”).

In fact, on your toothpick you can find the lead article from today’s *New York Times*. With enough squinting, measuring, and translating.

By producing the toothpick, they put the message there and thus they gave you NYT access, even though you are not a paid subscriber. You simply need to know how to stare (and translate), or in other words how to prompt.

So let’s sue the toothpick company! (Cowen, 2023)

Cowen is an economist, not a lawyer or computer scientist. But similar claims have been made by legal schol-

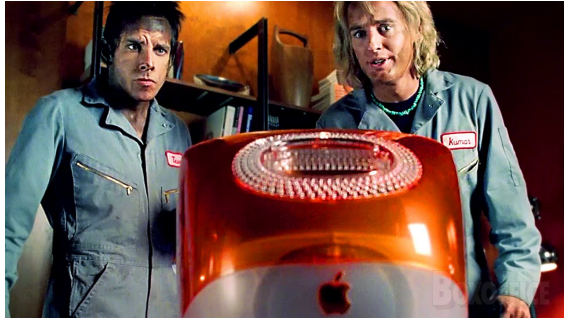


Figure 2. The files are in the computer.

ars (Bracha, 2024; Samuelson, 2024; Sprigman, 2024).

In this view, memorization in generative AI is inherently a phenomenon that takes place at *generation time*: when a user prompts a system and the system responds with an output. The model itself only learns abstracted features of training data, and represents those features in an extremely different and often uninterpretable way. Only when the model is prompted by a user in a suitably targeted (i.e., “adversarial”, “not normal”) way does a memorized output emerge. Thus, a generative-AI system is a general-purpose tool that some users may use to produce infringing outputs, but other users will not (Lee et al., 2023; Samuelson, 2023; Sag, 2023).

This view treats the machine-learned model(s) at the heart of a generative-AI system as a black box: something that receives training data as an input and is then capable of behaving in certain ways. But it refuses to consider what happens inside the box — the specifics of *how* statistical learning about the training data enables those behaviors. It avoids engaging with the actual representation of information about training data in a model’s parameters.

This way of thinking about memorization has significant copyright consequences. It suggests that memorization is primarily about *prompting* rather than *training*. Outputs may contain infringing expression, but the model that generates them does not. The model itself is a neutral tool, equally good at producing infringing and non-infringing outputs. Users bear most or all of the responsibility for misusing an AI system to elicit memorized content, and the creators of the system bear little or none.

With respect, we believe that this approach to memorization misdescribes how generative-AI systems work. (See Figure 2.) If a generative-AI system memorizes its training data, the training data is *in the model*. This should be unsurprising. Models are not inert tools that have no relationship with their training data. The power of a model is precisely that it encodes relevant features of the training data in a way that enables prompting to generate outputs that are based

on the training data. That is what generative AI *is*; that is what makes generative AI so powerful. All useful models learn something about their training data. Memorization is simply a difference in degree: it is an encoded feature *in the model*; whether it is a desired feature or not is another matter entirely.

It follows that memorization in generative AI cannot be neatly confined to generation time, to adversarial users, and to generation-time guardrails. If a generative-AI system has memorized copyrighted works, the memorized aspects of those works are present *in the model itself*, not just in the generated outputs. It can (with certain probability) generate near-verbatim copies of those works *on demand*, not just for users who have a suitably nefarious intent. And the system’s creator can limit output infringement by *changing the model*, not just by putting guardrails around the model (gpt, 2023).

We take no position on what the most appropriate copyright regimes for generative-AI systems should be, and we express no opinion on how pending copyright lawsuits should be decided. Our goal is merely to describe how these systems work so that copyright scholars can develop their theories of generative AI on a firm technical foundation. We seek clarity, precision, and technical accuracy.

If accepted, the camera-ready version of this paper will precede in two additional parts. First, we will provide a brief background on how generative-AI models work, and the supply chains within which they are embedded (Lee et al., 2023; 2024). Second, in the heart of the piece, we will describe how to think clearly about memorization in generative-AI systems, and show how several common arguments about copyright and generative AI are built on a mistaken view of how memorization happens. We will then offer a brief conclusion, with some historical reflections on scholarship at the intersection of technology and law (Newell, 1986).

References

- GPT-4 System Card. Technical report, OpenAI, March 2023. URL <https://cdn.openai.com/papers/gpt-4-system-card.pdf>.
- Oren Bracha. The Work of Copyright in the Age of Machine Production, January 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4581738.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

- USENIX Association, August 2021. ISBN 978-1-939133-24-3.
- Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models, 2023a.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. In *International Conference on Learning Representations*, 2023b.
- Chat GPT Is Eating the World, 2024. URL <https://chatgptiseatingtheworld.com>.
- Tyler Cowen. Toothpick producers violate NYT copyright, 2023. URL <https://marginalrevolution.com/marginalrevolution/2023/12/toothpick-producers-violate-nyt-copyright.html>.
- Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin’ ’Bout AI Generation: Copyright and the Generative-AI Supply Chain. *arXiv preprint arXiv:2309.08133*, 2023.
- Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin’ ’Bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version). In *Proceedings of the Symposium on Computer Science and Law, CSLAW ’24*, page 48–63, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703331. doi: 10.1145/3614407.3643696. URL <https://doi.org/10.1145/3614407.3643696>.
- Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable Extraction of Training Data from (Production) Language Models. *arXiv preprint arXiv:2311.17035*, 2023.
- New York Times Company v. Microsoft. 2:24-cv-00711 (C.D. Cal.).
- Allen Newell. Response: The Models Are Broken; The Models Are Broken. *University of Pittsburgh Law Review*, 47, 1986.
- OpenAI. ChatGPT: Optimizing Language Models for Dialogue, 2022. URL <https://web.archive.org/web/20221130180912/https://openai.com/blog/chatgpt/>.
- OpenAI. DALL-E 3 is now available in ChatGPT Plus and Enterprise, October 2023. URL <https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise>.
- Matthew Sag. Copyright Safety for Generative AI. *Houston Law Review*, 2023. Forthcoming.
- Pamela Samuelson. Generative AI meets copyright. *Science*, 381(6654):158–161, 2023.
- Pamela Samuelson. How to Think About Remedies in the Generative AI Copyright Cases. *Lawfare*, February 2024. URL <https://www.lawfaremedia.org/article/how-to-think-about-remedies-in-the-generative-ai-copyright-cases>.
- Christopher J. Sprigman. Upsetting Conventional Wisdom of Copyright Scholarship in the Age of AI. *Jotwell*, March 2024. URL <https://ip.jotwell.com/upsetting-conventional-wisdom-of-copyright-scholarship-in-the-age-of-ai/>.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C. Mozer, and Yoram Singer. Identity Crisis: Memorization and Generalization Under Extreme Overparameterization. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=B116y0VFPr>.