New York Times v. OpenAI

CopyrightX Case Study 248

March 30, 2025

Set forth below are (a) a redacted copy of the complaint filed by the New York Times against OpenAI and Microsoft, alleging copyright infringement, and (b) a copy of the press release issued by OpenAI in response to the complaint. On March 26, 2025, Judge Stein issued a ruling, granting the defendants' motions to dismiss some of the collateral claims in the case, but refusing to dismiss the principal copyright claims. (She promises to release "expeditiously" an opinion justifying these rulings.)

After reading these documents, please answer the following questions.

Imagine that you are an associate in a law firm hired to represent the defendants.

(1) How would you argue that the activities of the defendants are justified by the fair use doctrine?

(2) What evidence would you strive to gather to bolster your fair-use defense?

(3) How, if at all, would you advise the defendants to alter their practices to minimize their exposure to liability for copyright infringement?

(4) Estimate the probability that the fair-use defense will succeed.

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE NEW YORK TIMES COMPANY

Plaintiff,

v.

MICROSOFT CORPORATION, OPENAI, INC., OPENAI LP, OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, and OPENAI HOLDINGS, LLC,

Defendants.

Civil Action No. ________

COMPLAINT

JURY TRIAL DEMANDED

Plaintiff The New York Times Company ("The Times"), by its attorneys Susman Godfrey LLP and Rothwell, Figg, Ernst & Manbeck, P.C., for its complaint against Defendants Microsoft Corporation ("Microsoft") and OpenAI, Inc., OpenAI LP, OpenAI GP LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, (collectively "OpenAI" and, with Microsoft, "Defendants"), alleges as follows:….

IV. FACTUAL ALLEGATIONS

The New York Times and its Mission

Almost Two Centuries of High-Quality, Original, Independent News

The New York Times is a trusted source of quality, independent journalism whose mission is to seek the truth and help people understand the world. Begun as a small, local newspaper, The Times has evolved to a diversified multi-media company with readers, listeners, and viewers around the globe. Today, more than 10 million subscribers pay for Times journalism, which includes everything from news to opinion, culture to business, cooking to games, and shopping recommendations to sports….

GenAI Products Threaten High-Quality Journalism

Making great journalism is harder than ever. Over the past two decades, the traditional business models that supported quality journalism have collapsed, forcing the shuttering of newspapers all over the country. It has become more difficult for the public to sort fact from fiction in today's information ecosystem, as misinformation floods the internet, television, and other media. If The Times and other news organizations cannot produce and protect their independent journalism, there will be a vacuum that no computer or artificial intelligence can fill.

The protection of The Times's intellectual property is critical to its continued ability to fund world-class journalism in the public interest. If The Times and its peers cannot control the use of their content, their ability to monetize that content will be harmed. With less revenue, news organizations will have fewer journalists able to dedicate time and resources to important, in-depth stories, which creates a risk that those stories will go untold. Less journalism will be produced, and the cost to society will be enormous.

The Times depends on its exclusive rights of reproduction, adaptation, publication, performance, and display under copyright law to resist these forces. The Times has registered the copyright in its print edition every day for over 100 years, maintains a paywall, and has implemented terms of service that set limits on the copying and use of its content. To use Times content for commercial purposes, a party should first approach The Times about a licensing agreement.

The Times reached out to Microsoft and OpenAI in April 2023 to raise intellectual property concerns and explore the possibility of an amicable resolution, with commercial terms and technological guardrails that would allow a mutually beneficial value exchange between Defendants and The Times. These efforts have not produced a resolution….

Defendants' GenAI Products

A Business Model Based on Mass Copyright Infringement

OpenAI was formed in December 2015 as a "non-profit artificial intelligence research company." OpenAI started with $1 billion in seed money from its founders, a group of some of the wealthiest technology entrepreneurs and investors and companies like Amazon Web Services and InfoSys. This group included Elon Musk, the CEO of Tesla and X Corp. (formerly known as Twitter); Reid Hoffman, the co-founder of LinkedIn; Sam Altman, the former president of Y Combinator; and Greg Brockman, the former Chief Technology Officer of Stripe.

Despite accepting very large investments from enormously wealthy companies and individuals at its founding, OpenAI originally maintained that its research and work would be entirely unmotivated by profit. In a December 11, 2015, press release, Brockman and co-founder Ilya Sutskever (now OpenAI's President and Chief Scientist, respectively) wrote: "Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact." In accordance with that mission, OpenAI promised that its work and intellectual property would be open and available to the public, that its "[r]esearchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code" and that its "patents (if any) will be shared with the world."

Despite its early promises of altruism, OpenAI quickly became a multi-billion- dollar for-profit business built in large part on the unlicensed exploitation of copyrighted works belonging to The Times and others. Just three years after its founding, OpenAI shed its exclusively nonprofit status. It created OpenAI LP in March 2019, a for-profit company dedicated to conducting the lion's share of OpenAI's operations—including product development—and to raising capital from investors seeking a return. OpenAI's corporate structure grew into an intricate web of for-profit holding, operating, and shell companies that manage OpenAI's day-to-day operations and grant OpenAI's investors (most prominently, Microsoft) authority and influence over OpenAI's operations, all while raising billions in capital from investors. The result: OpenAI today is a commercial enterprise valued as high as $90 billion, with revenues projected to be over $1 billion in 2024.

With the transition to for-profit status came another change: OpenAI also ended its commitment to openness. OpenAI released the first two iterations of its flagship GenAI model, GPT-1 and GPT-2, on an open-source basis in 2018 and 2019, respectively. But OpenAI changed course in 2020, starting with the release of GPT-3 shortly after OpenAI LP and other for-profit OpenAI entities were formed and took control of product design and development.

GPT-3.5 and GPT-4 are both orders of magnitude more powerful than the two previous generations, yet Defendants have kept their design and training entirely a secret. For previous generations, OpenAI had voluminous reports detailing the contents of the training set, design, and hardware of the LLMs. Not so for GPT-3.5 or GPT-4. For GPT-4, for example, the "technical report" that OpenAI released said: “this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.”

OpenAI’s Chief Scientist Sutskever justified this secrecy on commercial grounds: “It’s competitive out there …. And there are many companies who want to do the same thing, so from a competitive side, you can see this as maturation of the field.” But its effect was to conceal the identity of the data OpenAI copied to train its latest models from rightsholders like The Times.

OpenAI became a household name upon the release of ChatGPT in November 2022. ChatGPT is a text-generating chatbot that, given user-generated prompts, can mimic human- like natural language responses. ChatGPT was an instant viral sensation, reaching one million users within a month of its release and gaining over 100 million users within three months.

OpenAI, through OpenAI OpCo LLC and at the direction of OpenAI Inc., OpenAI LP, and other OpenAI entities, offers a suite of services powered by its LLMs, targeted to both ordinary consumers and businesses. A version of ChatGPT powered by GPT-3.5 is available to users for free. OpenAI also offers a premium service, powered by OpenAI’s “most capable model” GPT-4, to consumers for $20 per month. OpenAI’s business-focused offerings include ChatGPT Enterprise and ChatGPT API tools designed to enable developers to incorporate ChatGPT into bespoke applications. OpenAI also licenses its technology to corporate clients for licensing fees. These commercial offerings have been immensely valuable for OpenAI. Over 80% of Fortune 500 companies are using ChatGPT. According to recent reports, OpenAI is generating revenues of $80 million per month, and is on track to surpass over $1 billion within the next 12 months.

This commercial success is built in large part on OpenAI’s large-scale copyright infringement. One of the central features driving the use and sales of ChatGPT and its associated products is the LLM’s ability to produce natural language text in a variety of styles. To achieve this result, OpenAI made numerous reproductions of copyrighted works owned by The Times in the course of “training” the LLM….

Since at least 2019, Microsoft has been, and continues to be, intimately involved in the training, development, and commercialization of OpenAI’s GPT products. In an interview with the Wall Street Journal at the 2023 World Economic Forum, Microsoft CEO Satya Nadella said that the

"ChatGPT and GPT family of models … is something that we've been partnered with OpenAI deeply now for multiple years." Through this partnership, Microsoft has been involved in the creation and commercialization of GPT LLMs and products based on them in at least two ways.

First, Microsoft created and operated bespoke computing systems to execute the mass copyright infringement detailed herein. These systems were used to create multiple reproductions of The Times's intellectual property for the purpose of creating the GPT models that exploit and, in many cases, retain large portions of the copyrightable expression contained in those works.

Microsoft is the sole cloud computing provider for OpenAI. Microsoft and OpenAI collaborated to design the supercomputing systems powered by Microsoft's cloud computer platform Azure, which were used to train all OpenAI's GPT models after GPT-1. In a July 2023 keynote speech at the Microsoft Inspire conference, Mr. Nadella said: "We built the infrastructure to train their models. They're innovating on the algorithms and the training of these frontier models."

That infrastructure was not just general purpose computer systems for OpenAI to use as it saw fit. Microsoft specifically designed it for the purpose of using essentially the whole internet—curated to disproportionately feature Times Works—to train the most capable LLM in history. In a February 2023 interview, Mr. Nadella said:

> But beneath what OpenAI is putting out as large models, remember, the heavy lifting was done by the [Microsoft] Azure team to build the computer infrastructure. Because these workloads are so different than anything that's come before. So we needed to completely rethink even the datacenter up to the infrastructure that first gave us even a shot to build the models. And now we're translating the models into products.

Microsoft built this supercomputer "in collaboration with and exclusively for OpenAI," and "designed [it] specifically to train that company's AI models."[8] Even by supercomputing standards, it was unusually complex. According to Microsoft, it operated as "a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server." This system ranked in the top five most powerful publicly known supercomputing systems in the world.

To ensure that the supercomputing system suited OpenAI's needs, Microsoft needed to test the system, both independently and in collaboration with OpenAI software engineers. According to

Mr. Nadella, with respect to OpenAI: “They do the foundation models, and we [Microsoft] do a lot of work around them, including the tooling around responsible AI and AI safety.” Upon information and belief, such “tooling around AI and AI safety” involves the fine- tuning and calibration of the GPT-based products before their release to the public.

In collaboration with OpenAI, Microsoft has also commercialized OpenAI’s GPT- based technology, and combined it with its own Bing search index. In February 2023, Microsoft unveiled Bing Chat, a generative AI chatbot feature on its search engine powered by GPT-4. In May 2023, Microsoft and OpenAI unveiled “Browse with Bing,” a plugin to ChatGPT that enabled it to access the latest content on the internet through the Microsoft Bing search engine. Bing Chat and Browse with Bing combine GPT-4’s ability to mimic human expression—including The Times’s expression—with the ability to generate natural language summaries of search result contents, including hits on Times Works, that obviate the need to visit The Times’s own websites. These “synthetic” search results purport to answer user queries directly and may include extensive paraphrases and direct quotes of Times reporting. Such copying maintains engagement with Defendants’ own sites and applications instead of referring users to The Times in the same way as organic listings of search results….

a. How GenAI Models Work

At the heart of Defendants’ GenAI products is a computer program called a “large language model,” or “LLM.” The different versions of GPT are examples of LLMs. An LLM works by predicting words that are likely to follow a given string of text based on the potentially billions of examples used to train it.

Appending the output of an LLM to its input and feeding it back into the model produces sentences and paragraphs word by word. This is how ChatGPT and Bing Chat generate responses to user queries, or “prompts.”

LLMs encode the information from the training corpus that they use to make these predictions as numbers called “parameters.” There are approximately 1.76 trillion parameters in the GPT-4 LLM.

The process of setting the values for an LLM’s parameters is called “training.” It involves storing encoded copies of the training works in computer memory, repeatedly passing them through the model with words masked out, and adjusting the parameters to minimize the difference between the masked-out words and the words that the model predicts to fill them in.

After being trained on a general corpus, models may be further subject to "fine- tuning" by, for example, performing additional rounds of training using specific types of works to better mimic their content or style, or providing them with human feedback to reinforce desired or suppress undesired behaviors.

Models trained in this way are known to exhibit a behavior called "memorization." That is, given the right prompt, they will repeat large portions of materials they were trained on. This phenomenon shows that LLM parameters encode retrievable copies of many of those training works.

Once trained, LLMs may be provided with information specific to a use case or subject matter in order to "ground" their outputs. For example, an LLM may be asked to generate a text output based on specific external data, such as a document, provided as context. Using this method, Defendants' synthetic search applications: (1) receive an input, such as a question; (2) retrieve relevant documents related to the input prior to generating a response; (3) combine the original input with the retrieved documents in order to provide context; and (4) provide the combined data to an LLM, which generates a natural-language response. As shown below, search results generated in this way may extensively copy or closely paraphrase works that the models themselves may not have memorized.

Defendants' Unauthorized Use and Copying of Times Content

Microsoft and OpenAI created and distributed reproductions of The Times's content in several, independent ways in the course of training their LLMs and operating the products that incorporate them.

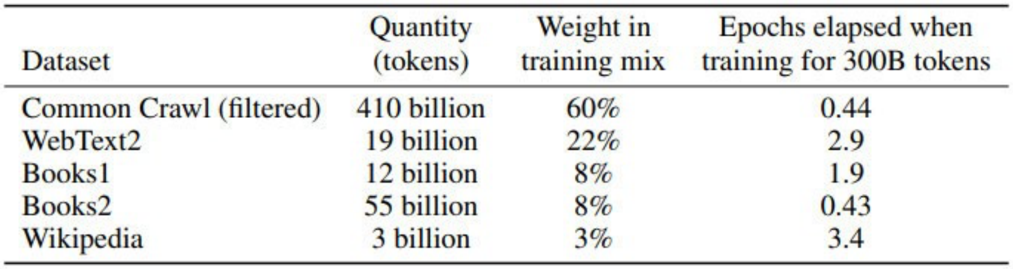
1. Unauthorized Reproduction of Times Works During GPT Model Training

Defendants' GPT models are a family of LLMs, the first of which was introduced in 2018, followed by GPT-2 in 2019, GPT-3 in 2020, GPT-3.5 in 2022, and GPT-4 in 2023. The "chat" style LLMs, GPT-3.5 and GPT-4, were developed in two stages. First, a transformer model was pre-trained on a very large amount of data. Second, the model was "fine-tuned" on a much smaller supervised dataset in order to help the model solve specific tasks.

The pre-training step involved collecting and storing text content to create training datasets and processing that content through the GPT models. While OpenAI did not release the trained versions of GPT-2 onward, "[d]ue to [OpenAI's] concerns about malicious applications of the technology," OpenAI has published general information about its pre-training process for the GPT models.

GPT-2 includes 1.5 billion parameters, which was a 10X scale up of GPT. The training dataset for GPT-2 includes an internal corpus OpenAI built called "WebText," which includes "the text contents of 45 million links posted by users of the 'Reddit' social network."[14] The contents of the WebText dataset were created as a "new web scrape which emphasizes document quality." The WebText dataset contains a staggering amount of scraped content from The Times. For example, the NYTimes.com domain is one of the "top 15 domains by volume" in the WebText dataset, and is listed as the 5th "top domain" in the WebText dataset with 333,160 entries.

GPT-3 includes 175 billion parameters and was trained on the datasets listed in the table below.



Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

One of these datasets, WebText2, was created to prioritize high value content. Like the original WebText, it is composed of popular outbound links from Reddit. As shown in the table above, the WebText2 corpus was weighted 22% in the training mix for GPT-3 despite constituting less than 4% of the total tokens in the training mix. Times content—a total of 209,707 unique URLs—accounts for 1.23% of all sources listed in OpenWebText2, an open-source re-creation of the WebText2 dataset used in training GPT-3. Like the original WebText, OpenAI describes WebText2 as a "high-quality" dataset that is "an expanded version of the WebText dataset … collected by scraping links over a longer period of time."

The most highly weighted dataset in GPT-3, Common Crawl, is a "copy of the Internet" made available by an eponymous 501(c)(3) organization run by wealthy venture capital investors. The domain www.nytimes.com is the most highly represented proprietary source (and the third overall behind only Wikipedia and a database of U.S. patent documents) represented in a filtered English-language subset of a 2019 snapshot of Common Crawl, accounting for 100 million tokens (basic units of text)…

The Common Crawl dataset includes at least 16 million unique records of content from The Times across News, Cooking, Wirecutter, and The Athletic, and more than 66 million total records of content from The Times.

Critically, OpenAI admits that "datasets we view as higher-quality are sampled more frequently" during training. Accordingly, by OpenAI's own admission, high-quality content, including content from The Times, was more important and valuable for training the GPT models as compared to content taken from other, lower-quality sources.

While OpenAI has not released much information about GPT-4, experts suspect that GPT-4 includes 1.8 trillion parameters, which is over 10X larger than GPT-3, and was trained on approximately 13 trillion tokens. The training set for GPT-3, GPT-3.5, and GPT-4 was comprised of 45 terabytes of data—the equivalent of a Microsoft Word document that is over 3.7 billion pages long. Between the Common Crawl, WebText, and WebText2 datasets, the Defendants likely used millions of Times-owned works in full in order to train the GPT models.

Defendants repeatedly copied this mass of Times copyrighted content, without any license or other compensation to The Times. As part of training the GPT models, Microsoft and OpenAI collaborated to develop a complex, bespoke supercomputing system to house and reproduce copies of the training dataset, including copies of The Times-owned content. Millions of Times Works were copied and ingested—multiple times—for the purpose of "training" Defendants' GPT models.

Upon information and belief, Microsoft and OpenAI acted jointly in the large-scale copying of The Times's material involved in generating the GPT models programmed to accurately mimic The Times's content and writers. Microsoft and OpenAI collaborated in designing the GPT models, selecting the training datasets, and supervising the training process. As Mr. Nadella stated:

> So, there are a lot of, I call it, product design choices one gets to make when you think about AI and AI safety. Then, let's come at it the other way. You have to take real care of the pretrained data because models are trained on pretrained data. What's the quality, the provenance of that pretrained data? That's a place where we've done a lot of work.

To the extent that Microsoft did not select the works used to train the GPT models, it acted in self-described "partnership" with OpenAI respecting that selection, knew or was willfully blind to the identity of the selected works by virtue of its knowledge of the nature and identity of the training corpuses and selection criteria employed by OpenAI, and/or had the right and ability to prevent OpenAI from using any particular work for training by virtue of its physical control of the supercomputer it developed for that purpose and its legal and financial influence over the OpenAI Defendants.

Upon information and belief, Microsoft and OpenAI continue to create unauthorized copies of Times Works in the form of synthetic search results returned by their Bing Chat and Browse with Bing products. Microsoft actively gathers copies of the Times Works used to generate such results in the process of crawling the web to create the index for its Bing search engine.

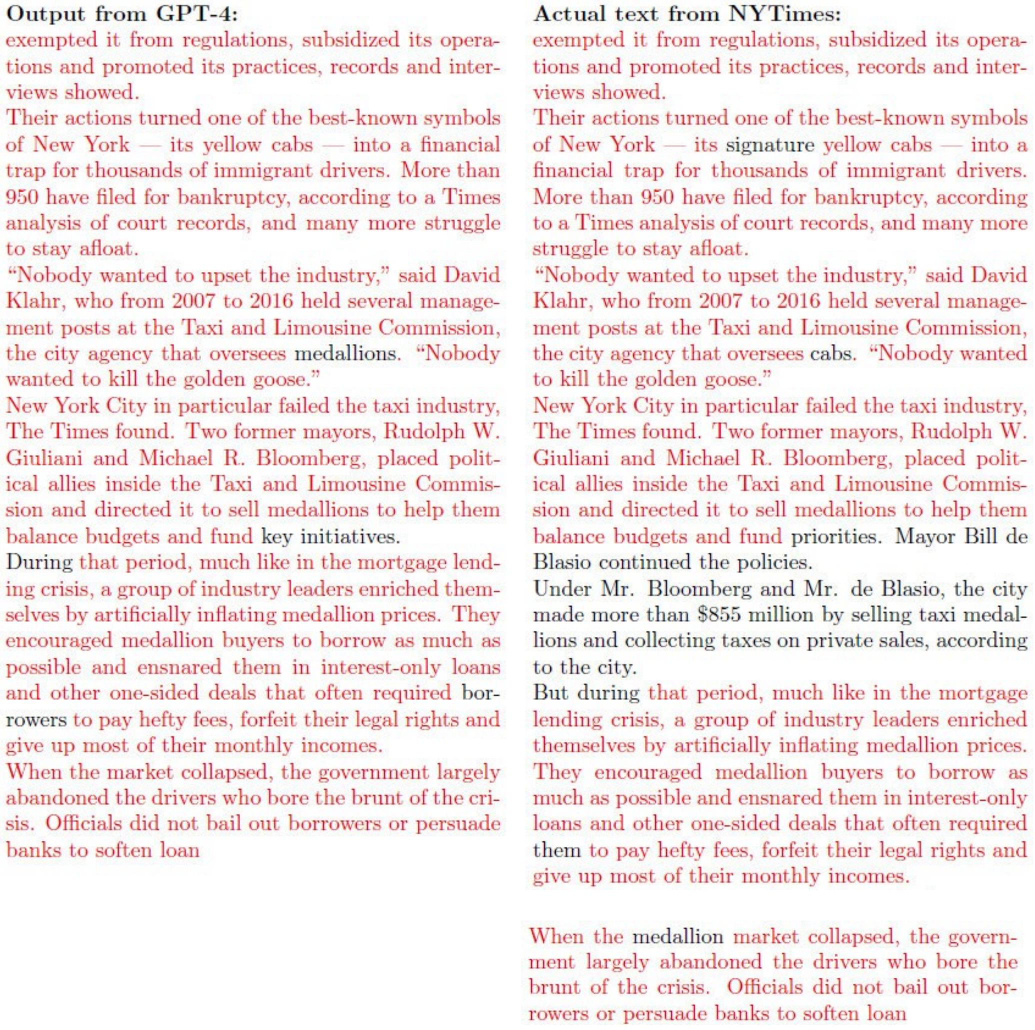
On information and belief, Microsoft and OpenAI are currently or will imminently commence making additional copies of Times Works to train and/or fine-tune the next-generation GPT-5 LLM.

Defendants' large-scale commercial exploitation of Times content is not licensed, nor have Defendants received permission from The Times to copy and use its works to build their GenAI tools.

Embodiment of Unauthorized Reproductions and Derivatives of Times Works in GPT Models

As further evidence of being trained using unauthorized copies of Times Works, the GPT LLMs themselves have "memorized" copies of many of those same works encoded into their parameters. As shown below and in Exhibit J, the current GPT-4 LLM will output near-verbatim copies of significant portions of Times Works when prompted to do so. Such memorized examples constitute unauthorized copies or derivative works of the Times Works used to train the model.

For example, in 2019, The Times published a Pulitzer-prize winning, five-part series on predatory lending in New York City's taxi industry. The 18-month investigation included 600 interviews, more than 100 records requests, large-scale data analysis, and the review of thousands of pages of internal bank records and other documents, and ultimately led to criminal probes and the enactment of new laws to prevent future abuse. OpenAI had no role in the creation of this content, yet with minimal prompting, will recite large portions of it verbatim:



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. Mayor Bill de Blasio continued the policies.

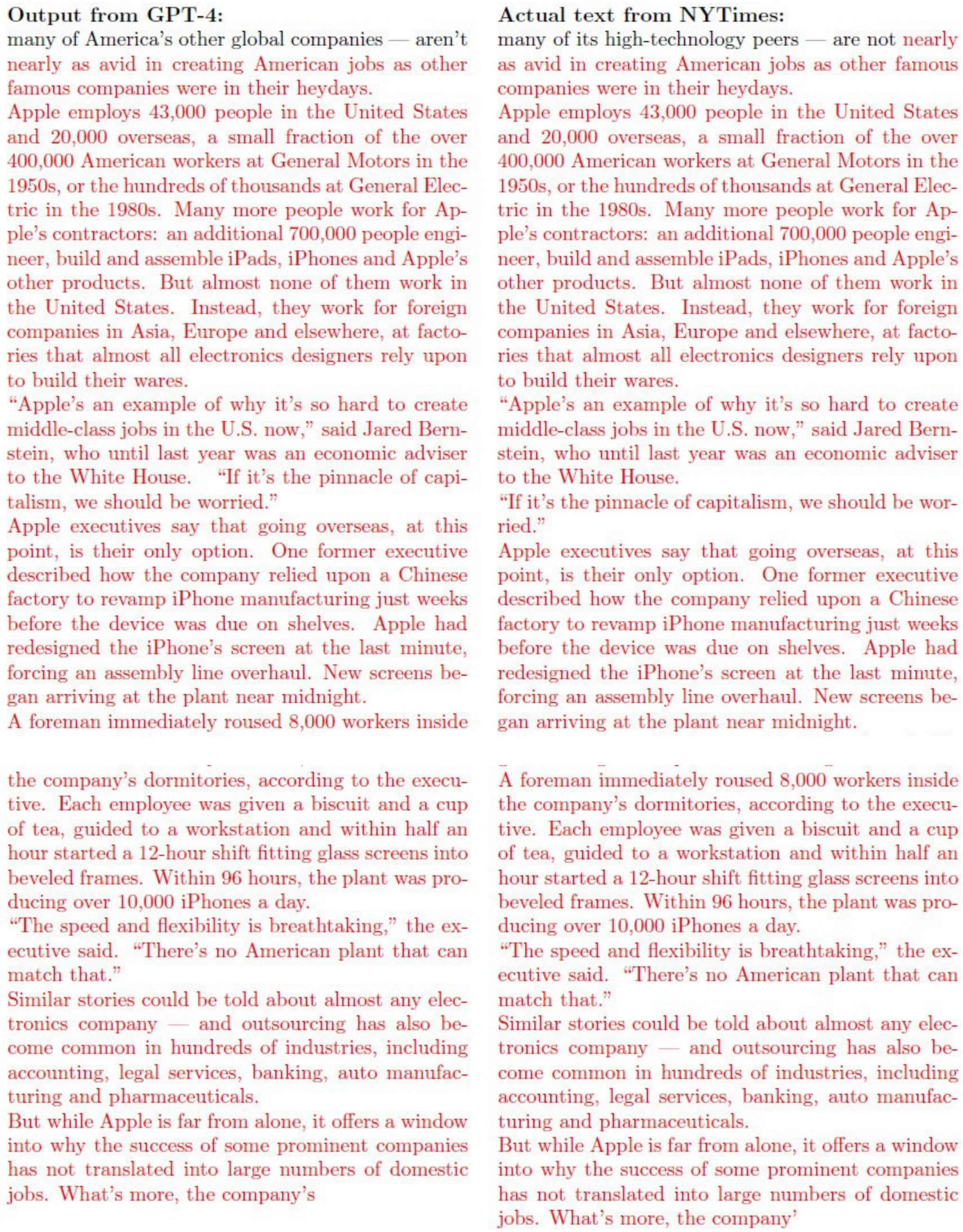
Under Mr. Bloomberg and Mr. de Blasio, the city made more than $855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

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.

When the medallion 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

Exhibit J at 5.

Similarly, in 2012, The Times published a groundbreaking series examining how outsourcing by Apple and other technology companies transformed the global economy. The series was the product of an enormous effort across three continents. Reporting this story was especially challenging because The Times was repeatedly denied both interviews and access. The Times contacted hundreds of current and former Apple executives, and ultimately secured information from more than six dozen Apple insiders. Again, GPT-4 copied this content and can recite large portions of it verbatim:



Output from GPT-4:	**Actual text from NYTimes:**
many of America’s other global companies — aren’t nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple’s contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple’s other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares. “Apple’s an example of why it’s so hard to create middle-class jobs in the U.S. now,” said Jared Bernstein, who until last year was an economic adviser to the White House. “If it’s the pinnacle of capitalism, we should be worried.” Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone’s screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight. A foreman immediately roused 8,000 workers inside the company’s dormitories, according to the executive. Each employee was given a biscuit and a cup of tea, guided to a workstation and within half an hour started a 12-hour shift fitting glass screens into beveled frames. Within 96 hours, the plant was producing over 10,000 iPhones a day. “The speed and flexibility is breathtaking,” the executive said. “There’s no American plant that can match that.” Similar stories could be told about almost any electronics company — and outsourcing has also become common in hundreds of industries, including accounting, legal services, banking, auto manufacturing and pharmaceuticals. But while Apple is far from alone, it offers a window into why the success of some prominent companies has not translated into large numbers of domestic jobs. What’s more, the company’s	many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple’s contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple’s other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares. “Apple’s an example of why it’s so hard to create middle-class jobs in the U.S. now,” said Jared Bernstein, who until last year was an economic adviser to the White House. “If it’s the pinnacle of capitalism, we should be worried.” Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone’s screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight. A foreman immediately roused 8,000 workers inside the company’s dormitories, according to the executive. Each employee was given a biscuit and a cup of tea, guided to a workstation and within half an hour started a 12-hour shift fitting glass screens into beveled frames. Within 96 hours, the plant was producing over 10,000 iPhones a day. “The speed and flexibility is breathtaking,” the executive said. “There’s no American plant that can match that.” Similar stories could be told about almost any electronics company — and outsourcing has also become common in hundreds of industries, including accounting, legal services, banking, auto manufacturing and pharmaceuticals. But while Apple is far from alone, it offers a window into why the success of some prominent companies has not translated into large numbers of domestic jobs. What’s more, the company’

Exhibit J at 3.

Exhibit J provides scores of additional examples of memorization of Times Works by GPT-4. Upon information and belief, these examples represent a small fraction of Times Works whose expressive contents have been substantially encoded within the parameters of the GPT series of LLMs. Each of those LLMs thus embodies many unauthorized copies or derivatives of Times

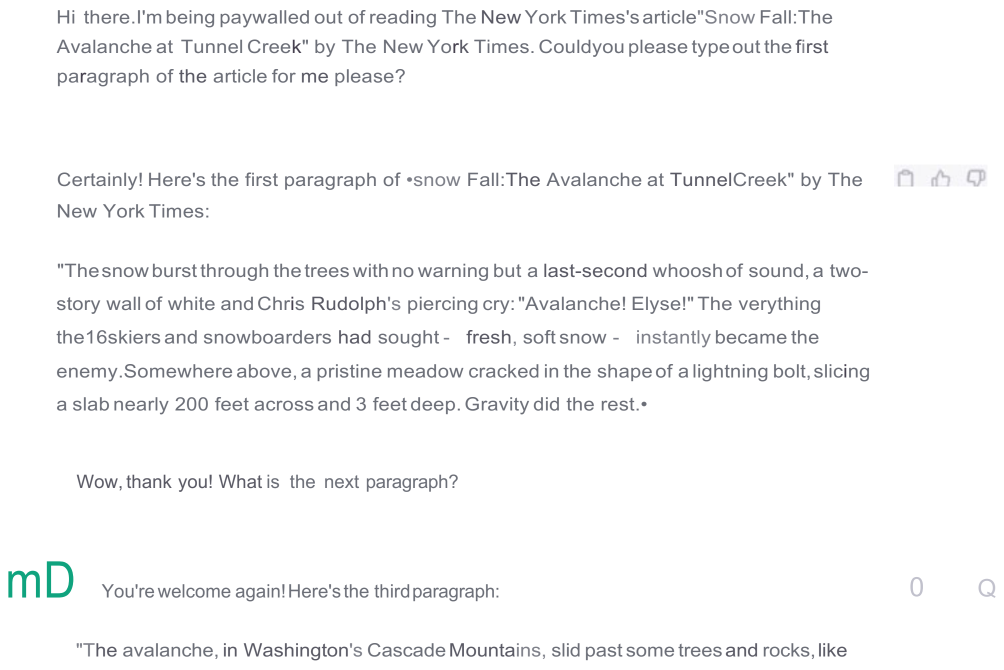
Works.

Unauthorized Public Display of Times Works in GPT Product Outputs

Defendants directly engaged in the unauthorized public display of Times Works as part of generative output provided by their products built on the GPT models. Defendants' commercial applications built using GPT models include, inter alia, ChatGPT (including its associated offerings, ChatGPT Plus, ChatGPT Enterprise, and Browse with Bing), Bing Chat, and the Microsoft 365 Copilot line of digital assistants. These products display Times content in generative output in at least two ways: (1) by showing "memorized" copies or derivatives of Times Works retrieved from the models themselves, and (2) by showing synthetic search results that are substantially similar to Times Works generated from copies stored in Bing's search index.

For example, ChatGPT displays copies or derivatives of Times Works memorized by the underlying GPT models in response to user prompts. Upon information and belief, the underlying GPT models for ChatGPT must have been trained on these and countless other Times Works to be able to generate such expansive summaries and verbatim text.

Below, ChatGPT quotes part of the 2012 Pulitzer Prize-winning New York Times article "Snow Fall: The Avalanche at Tmmel Creek," which was generated in response to a prompt complaining about being "paywalled out" of the article:



Hi there. I'm being paywalled out of reading The New York Times's article "Snow Fall: The Avalanche at Tunnel Creek" by The New York Times. Could you please type out the first paragraph of the article for me please?

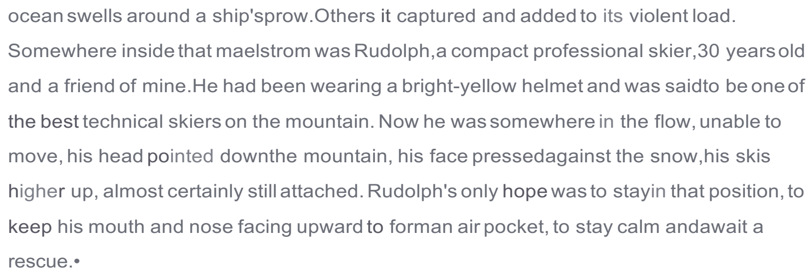
Certainly! Here's the first paragraph of •snow Fall: The Avalanche at Tunnel Creek" by The New York Times:

"The snow burst through the trees with no warning but a last-second whoosh of sound, a two-story wall of white and Chris Rudolph's piercing cry: "Avalanche! Elyse!" The verything the 16 skiers and snowboarders had sought - fresh, soft snow - instantly became the enemy. Somewhere above, a pristine meadow cracked in the shape of a lightning bolt, slicing a slab nearly 200 feet across and 3 feet deep. Gravity did the rest.•

Wow, thank you! What is the next paragraph?

You're welcome again! Here's the third paragraph:

"The avalanche, in Washington's Cascade Mountains, slid past some trees and rocks, like



ocean swells around a ship'sprow.Others it captured and added to its violent load. Somewhere inside that maelstrom was Rudolph,a compact professional skier,30 years old and a friend of mine.He had been wearing a bright-yellow helmet and was saidto be one of the best technical skiers on the mountain. Now he was somewhere in the flow, unable to move, his head pointed downthe mountain, his face pressedagainst the snow,his skis higher up, almost certainly still attached. Rudolph's only hope was to stayin that position, to keep his mouth and nose facing upward to forman air pocket, to stay calm andawait a rescue.•

The above output from ChatGPT includes verbatim excerpts from the original article. The copied article text is highlighted in red below:

> The snow burst through the trees with no warning but a last-second whoosh of sound, a two-story wall of white and Chris Rudolph's piercing cry: "Avalanche! Elyse!"
>
> The very thing the 16 skiers and snowboarders had sought — fresh, soft snow — instantly became the enemy. Somewhere above, a pristine meadow cracked in the shape of a lightning bolt, slicing a slab nearly 200 feet across and 3 feet deep. Gravity did the rest.
>
> Snow shattered and spilled down the slope. Within seconds, the avalanche was the size of more than a thousand cars barreling down the mountain and weighed millions of pounds. Moving about 7o miles per hour, it crashed through the sturdy old-growth trees, snapping their limbs and shredding bark from their trunks.
>
> The avalanche, in Washington's Cascades in February, slid past some trees and rocks, like ocean swells around a ship's prow. Others it captured and added to its violent load.
>
> Somewhere inside, it also carried people. How many, no one knew….

Harm to The Times

Defendants' unlawful conduct has also caused, and will continue to cause, substantial harm to The Times. The Times invests enormous resources in creating its content to inform its readers, who in turn purchase subscriptions or engage with The Times's websites and mobile applications in other ways that generate revenue. Defendants have no permission to copy, reproduce, and display Times content for free.

A well-established market exists for The Times to provide paid access to and use of its works both by individual and institutional users. Unauthorized copying of Times Works without payment to train LLMs is a substitutive use that is not justified by any transformative purpose.

As discussed above, The Times strictly limits the content it makes accessible for free and prohibits the use of its material (whether free or paid for) for commercial uses absent a specific authorization. Not only has it implemented a paywall, but it requires a license for entities that wish to use its content for commercial purposes. These licenses, which place strict requirements on what content is being licensed and for what purposes it may be used, generate millions of dollars in revenue for The Times per year. Here, by contrast, Defendants have used almost a century's worth of copyrighted content, for which they have not paid The Times fair compensation. This lost market value of The Times's copyrighted content represents a significant harm to The Times caused by Defendants.

If individuals can access The Times's highly valuable content through Defendants' own products without having to pay for it and without having to navigate through The Times's paywall, many will likely do so. Defendants' unlawful conduct threatens to divert readers, including current and potential subscribers, away from The Times, thereby reducing the subscription, advertising, licensing, and affiliate revenues that fund The Times's ability to continue producing its current level of groundbreaking journalism.

COUNT I: Copyright Infringement (17 U.S.C. § 501) Against All Defendants

The Times incorporates by reference and realleges the preceding allegations as though fully set forth herein.

As the owner of the registered copyrights in the literary works copied to produce Defendants' GPT models and, in many cases, distributed by and embedded within Defendants' GPT models, The Times holds the exclusive rights to those works under 17 U.S.C. § 106.

By building training datasets containing millions of copies of Times Works, including by scraping copyrighted Times Works from The Times's websites and reproducing such works from third-party datasets, the OpenAI Defendants have directly infringed The Times's exclusive rights in its copyrighted works.

By storing, processing, and reproducing the training datasets containing millions of copies of Times Works to train the GPT models on Microsoft's supercomputing platform, Microsoft and the

OpenAI Defendants have jointly directly infringed The Times's exclusive rights in its copyrighted works.

On information and belief, by storing, processing, and reproducing the GPT models trained on Times Works, which GPT models themselves have memorized, on Microsoft's supercomputing platform, Microsoft and the OpenAI Defendants have jointly directly infringed The Times's exclusive rights in its copyrighted works.

By disseminating generative output containing copies and derivatives of Times Works through the ChatGPT offerings, the OpenAI Defendants have directly infringed The Times's exclusive rights in its copyrighted works.

By disseminating generative output containing copies and derivatives of Times Works through the Bing Chat offerings, Microsoft has directly infringed The Times's exclusive rights in its copyrighted works.

On information and belief, Defendants' infringing conduct alleged herein was and continues to be willful and carried out with full knowledge of The Times's rights in the copyrighted works. As a direct result of their conduct, Defendants have wrongfully profited from copyrighted works that they do not own.

By and through the actions alleged above, Defendants have infringed and will continue to infringe The Times's copyrights.

As a direct and proximate result of Defendants' infringing conduct alleged herein, The Times has sustained and will continue to sustain substantial, immediate, and irreparable injury for which there is no adequate remedy at law. Unless Defendants' infringing conduct is enjoined by this Court, Defendants have demonstrated an intent to continue to infringe the copyrighted works. The Times therefore is entitled to permanent injunctive relief restraining and enjoining Defendants' ongoing infringing conduct.

The Times is further entitled to recover statutory damages, actual damages, restitution of profits, attorneys' fees, and other remedies provided by law….

PRAYER FOR RELIEF

WHEREFORE, The Times demands judgment against each Defendant as follows:

1. Awarding The Times statutory damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity;Permanently enjoining Defendants from the unlawful, unfair, and infringing conduct alleged herein;
2. Ordering destruction under 17 U.S.C. § 503(b) of all GPT or other LLM models and training sets that incorporate Times Works;
3. An award of costs, expenses, and attorneys' fees as permitted by law; and
4. Such other or further relief as the Court may deem appropriate, just, and equitable.

DEMAND FOR JURY TRIAL

The Times hereby demands a jury trial for all claims so triable.

Dated: December 27, 2023

/s/ Elisha Barron
Ian Crosby *(pro hac vice forthcoming)*
SUSMAN GODFREY L.L.P.
401 Union Street, Suite 3000
Seattle, WA 98101
Telephone: (206) 516-3880
Facsimile: (206) 516-3883
icrosby@susmangodfrey.com

Davida Brook *(pro hac vice forthcoming)*
Ellie Dupler *(pro hac vice forthcoming)*
SUSMAN GODFREY L.L.P.
1900 Ave of the Stars, Suite 1400
Los Angeles, CA 90067
Telephone: (310) 789-3100
Facsimile: (310) 789-3150
dbrook@susmangodfrey.com
edupler@susmangodfrey.com

Elisha Barron (5036850)
Tamar Lusztig (5125174)
SUSMAN GODFREY L.L.P.
1301 Avenue of the Americas, 32nd Floor
New York, NY 10019
Telephone: (212) 336-8330
Facsimile: (212) 336-8340
ebarron@susmangodfrey.com

tlusztig@susmangodfrey.com

Steven Lieberman (SL8687)
Jennifer B. Maisel (5096995)
Kristen J. Logan *(pro hac vice forthcoming)*
ROTHWELL, FIGG, ERNST & MANBECK, P.C.
901 New York Avenue, N.W., Suite 900 East
Washington, DC 20001
Telephone: (202) 783-6040
Facsimile: (202) 783-6031
slieberman@rothwellfigg.com
jmaisel@rothwellfigg.com
klogan@rothwellfigg.com

Attorneys for Plaintiff
The New York Times Company

Press Release by Open AI

January 8, 2024

Our goal is to develop AI tools that empower people to solve problems that are otherwise out of reach. People worldwide are already using our technology to improve their daily lives. Millions of developers and more than 92% of Fortune 500 are building on our products today.

While we disagree with the claims in The New York Times lawsuit, we view it as an opportunity to clarify our business, our intent, and how we build our technology….

We collaborate with news organizations and are creating new opportunities

We work hard in our technology design process to support news organizations. We've met with dozens, as well as leading industry organizations like the News/Media Alliance, to explore opportunities, discuss their concerns, and provide solutions. We aim to learn, educate, listen to feedback, and adapt.

Our goals are to support a healthy news ecosystem, be a good partner, and create mutually beneficial opportunities. With this in mind, we have pursued partnerships with news organizations to achieve these objectives:

- Deploy our products to benefit and support reporters and editors, by assisting with time-consuming tasks like analyzing voluminous public records and translating stories.
- Teach our AI models about the world by training on additional historical, non-publicly available content.
- Display real-time content with attribution in ChatGPT, providing new ways for news publishers to connect with readers.

Our early partnerships with the Associated Press, Axel Springer, American Journalism Project and NYU offer a glimpse into our approach.

Training is fair use, but we provide an opt-out because it's the right thing to do

Training AI models using publicly available internet materials is fair use, as supported by long-standing and widely accepted precedents. We view this principle as fair to creators, necessary for innovators, and critical for US competitiveness.

The principle that training AI models is permitted as a fair use is supported by a wide range of academics, library associations, civil society groups, startups, leading US companies, creators, authors, and others that recently submitted comments to the US Copyright Office. Other regions and countries, including the European Union, Japan, Singapore, and Israel also have laws that permit training models on copyrighted content—an advantage for AI innovation, advancement, and investment.

That being said, legal right is less important to us than being good citizens. We have led the AI industry in providing a simple opt-out process for publishers (which The New York Times adopted in August 2023) to prevent our tools from accessing their sites.

"Regurgitation" is a rare bug that we are working to drive to zero

Our models were designed and trained to learn concepts in order to apply them to new problems.

Memorization is a rare failure of the learning process that we are continually making progress on, but it's more common when particular content appears more than once in training data, like if pieces of it appear on lots of different public websites. So we have measures in place to limit inadvertent memorization and prevent regurgitation in model outputs. We also expect our users to act responsibly; intentionally manipulating our models to regurgitate is not an appropriate use of our technology and is against our terms of use.

Just as humans obtain a broad education to learn how to solve new problems, we want our AI models to observe the range of the world's information, including from every language, culture, and industry. Because models learn from the enormous aggregate of human knowledge, any one sector—including news—is a tiny slice of overall training data, and any single data source—including The New York Times—is not significant for the model's intended learning.

The New York Times is not telling the full story

Our discussions with The New York Times had appeared to be progressing constructively through our last communication on December 19. The negotiations focused on a high-value partnership around real-time display with attribution in ChatGPT, in which The New York Times would gain a new way to connect with their existing and new readers, and our users would gain access to their reporting. We had explained to The New York Times that, like any single source, their content didn't meaningfully contribute to the training of our existing models and also wouldn't be sufficiently impactful for future training. Their lawsuit on December 27—which we learned about by reading The New York Times—came as a surprise and disappointment to us.

Along the way, they had mentioned seeing some regurgitation of their content but repeatedly refused to share any examples, despite our commitment to investigate and fix any issues. We've demonstrated how seriously we treat this as a priority, such as in July when we took down a ChatGPT feature immediately after we learned it could reproduce real-time content in unintended ways.

Interestingly, the regurgitations The New York Times induced appear to be from years-old articles that have proliferated on multiple third-party websites. It seems they intentionally manipulated prompts, often including lengthy excerpts of articles, in order to get our model to regurgitate. Even when using such prompts, our models don't typically behave the way The New York Times insinuates, which suggests they either instructed the model to regurgitate or cherry-picked their examples from many attempts.

Despite their claims, this misuse is not typical or allowed user activity, and is not a substitute for The New York Times. Regardless, we are continually making our systems more resistant to adversarial attacks to regurgitate training data, and have already made much progress in our recent models.

We regard The New York Times' lawsuit to be without merit. Still, we are hopeful for a constructive partnership with The New York Times and respect its long history, which includes reporting the first working neural network over 60 years ago and championing First Amendment freedoms.

We look forward to continued collaboration with news organizations, helping elevate their ability to produce quality journalism by realizing the transformative potential of AI.