Allegations and Proposed Questions

Allegation	Proposed Question
Defendants copied and used millions of New York Times works to build large language models without permission.	Who authorized the use of New York Times content in any training datasets?
Defendants gave New York Times content particular emphasis during model training.	How was New York Times content prioritized or weighted in model training?
Defendants' tools can output verbatim recitations, close summaries, or style mimicry of New York Times content.	What processes cause model outputs to reproduce or closely summarize specific New York Times articles?
Defendants' tools sometimes attribute false information to the New York Times.	When did any system output attribute statements to the New York Times that were later determined to be false?
Defendants use Microsoft's Bing search index, which copies and categorizes New York Times content, to generate responses with verbatim excerpts and detailed summaries.	How does the Bing index retrieve and pass New York Times content to downstream generative systems?
Generated responses are longer and more detailed than traditional search snippets and are provided without Times authorization, harming relationships and revenue.	What internal assessments measured traffic or revenue impact from generative answers that included New York Times content?
OpenAI's GPT models memorized copies of New York Times works encoded in model parameters, enabling near-verbatim outputs.	How do the models retrieve memorized passages derived from New York Times sources when prompted?
GPT-4 produced near-verbatim text from specific New York Times investigations.	What prompts and contexts lead GPT-4 to reproduce text from identified New York Times investigations?
Internal exhibits show additional examples of memorization of New York Times works by GPT-4.	Where are internal test logs showing outputs that match New York Times passages verbatim or near-verbatim?
ChatGPT displayed verbatim excerpts of paywalled New York Times content in response to user prompts.	When did any product output verbatim text from paywalled New York Times articles?
Defendants directly engaged in unauthorized public display of New York Times works via outputs in ChatGPT, Bing Chat, and Microsoft 365 Copilot.	Which product features displayed New York Times content in generated answers?
Synthetic search combines GPT-4 with the Bing index to generate summaries of search results, including New York Times hits, reducing visits to Times sites.	How are summaries of New York Times pages constructed from indexed content during answer generation?
Microsoft has been intimately involved in training, development, and commercialization of OpenAI's GPT products since at least 2019.	Who at Microsoft oversaw collaboration with OpenAI on GPT training and productization?
Microsoft created and operated bespoke computing systems used to reproduce New York Times content for GPT training.	Where were training copies of New York Times content stored and processed within Microsoft infrastructure?
Microsoft is the sole cloud provider and co-designed the Azure supercomputer used to train GPT models after GPT-1.	How did Azure supercomputing resources handle datasets containing New York Times material?
The Azure supercomputer built for OpenAI had substantial compute resources (e.g., hundreds of thousands of CPU cores and thousands of GPUs).	What safeguards were implemented to control copying of New York Times content on the training cluster?
Microsoft and OpenAI collaborated on responsible AI and safety tooling, including fine-tuning and calibration.	How did any fine-tuning workflows use New York Times content or outputs referencing the Times?
Microsoft and OpenAI commercialized GPT technology and combined it with Bing search, launching products like Bing Chat and Browse with Bing.	What product integrations combined Bing index results with generative answers containing New York Times content?
Microsoft leadership stated it has all necessary IP rights and capabilities even without OpenAI.	What rights did Microsoft claim regarding data, models, or outputs that include New York Times material?
OpenAI transitioned from a nonprofit mission to a complex for profit structure, raising significant funding with Microsoft influence.	When did OpenAI entities begin operating for profit, and how did that change data acquisition practices?
OpenAI ended its commitment to openness starting with GPT 3, keeping model design and training details secret.	What internal policies governed disclosure of training data sources that included New York Times content?
GPT 2's WebText dataset (built from Reddit links) included NYTimes.com among the top domains.	What proportion of WebText entries were sourced from NYTimes.com?
GPT 3 training used WebText2 and OpenWebText2, which included a large number of New York Times URLs.	Which New York Times URLs from OpenWebText2 were included in GPT 3 training data?
Common Crawl contains extensive New York Times records that appeared in training data sources.	How many New York Times records from Common Crawl were ingested into any training pipelines?
Higher quality datasets are sampled more frequently during training, increasing representation of certain sources.	Why were sources classified as higher quality, and how did that classification affect sampling of New York Times content?
Microsoft and OpenAI jointly designed models, selected training datasets, and supervised the training process.	Who decided to include specific New York Times sources in the training mixes?
Defendants repeatedly copied New York Times content multiple times for training on bespoke supercomputing systems.	How many copies of each New York Times item were created or cached during training runs?
Defendants' products display New York Times content either from model memorization or from synthetic search built on the Bing index.	How does the system decide between retrieving memorized text and summarizing indexed New York Times pages?
Defendants' conduct threatens to divert readers and reduce subscription, advertising, licensing, and affiliate revenues for the Times.	What analyses forecast reader diversion or revenue loss linked to generative answers containing New York Times content?
The Times restricts free access, requires licenses for commercial use, and generates substantial revenue from such licenses.	What licenses exist—or were sought—for commercial reuse of New York Times content in AI training or products?
Defendants have no permission to copy, reproduce, or display New York Times content for free.	When did any party request, grant, or deny permission to use New York Times content for AI purposes?
Defendants' use of New York Times content has been extremely lucrative for them.	How were product or valuation gains tied to features that leverage New York Times content?
The Times objected and negotiated for months to seek fair value and responsible use, but talks did not resolve the dispute.	Who participated in negotiations with the New York Times regarding content use, and what terms were proposed?
Defendants planned or commenced making additional copies of New York Times works to train or fine tune GPT 5.	When were copies of New York Times content created or scheduled for use in GPT 5 training or fine tuning?