

**Lieff
Cabraser
Heimann &
Bernstein**
Attorneys at Law

SUSMAN GODFREY L.L.P.

A REGISTERED LIMITED LIABILITY PARTNERSHIP



May 6, 2024

Hon. Sidney H. Stein
Daniel Patrick Moynihan
United States Courthouse
500 Pearl St.
New York, NY 10007-1312

RE: *Authors Guild et al. v. OpenAI, Inc., et al.*, and *Alter et al. v. OpenAI Inc., et al.*, Nos. 1:23-cv-08292-SHS & 1:23-cv-10211-SHS

Dear Judge Stein:

Pursuant to Rule 2(G) of Your Honor's Individual Practices, Plaintiffs seek an informal discovery conference concerning OpenAI's refusal to produce (1) a subset of its responses to interrogatories served on it by the Federal Trade Commission (FTC); and (2) documents sufficient to show the identities of two former employees responsible for creating two compendiums of books used in OpenAI's training data, that have since been deleted. Each addresses the core factual question of the case: What data did OpenAI use to "train" its Large Language Models (LLMs)? There is minimal burden to OpenAI in producing these documents and it should be compelled to provide it.

FTC INTERROGATORY RESPONSES

Background. OpenAI has been under investigation since before this case was filed. Though those investigations differ in focus from this case, there is overlap. For example, the FTC's Civil Investigative Demand (CID) includes the following interrogatory: "Describe in Detail the data You have used . . . to train or otherwise develop each Large Language Model [that You have provided, offered or made available since June 1, 2020]." Ex. A at 5. This has obvious relevance to Plaintiffs' allegations that OpenAI used (and copied) Plaintiffs' works to train their LLMs.

On November 13, 2023, Plaintiffs served a Request for Production of "Documents that You . . . have already gathered or collected in order to submit them to any legislative or executive agency, committee, or other governmental entity in the United States that Concern or Relate To the allegations in the Complaint." Ex. B, RFP No. 2. On December 22, 2023, OpenAI responded that it would not produce any documents responsive to this request. The parties have met and conferred regarding this request for documents from the FTC investigation over the course of several months and, on March 26, 2024, OpenAI confirmed that it would produce *none* of the documents or interrogatory responses that it has provided to the FTC, and that the parties were at an impasse. While OpenAI confirmed that it served written responses to certain FTC interrogatories, it refuses to disclose which ones.

Hon. Sidney H. Stein
 May 6, 2024
 Page 2

Discussion. Plaintiffs seek OpenAI's responses to 21 of the FTC's 49 Interrogatories. Ex. A.¹

Relevance and Burden. This litigation is about whether OpenAI used Plaintiffs' works to train its LLM products and whether such use infringed Plaintiffs' copyrights. While the FTC's focus is on consumer privacy, its Interrogatories Nos. 1-21 seek basic information about OpenAI's corporate structure (Nos. 1-7),² LLM products (Nos. 8-11), efforts to market its LLM products (Nos. 13-14), and training and development of its LLMs (Nos. 15-21). The information requested includes a description of "each Large Language Model Product [] that You have provided, offered or made available" since June 1, 2020 (No. 9), the data OpenAI used to train each of those LLM products (No. 15), and the training and refinement processes for each LLM (Nos. 18-19). Ex. A at 2-17.

OpenAI's response to these interrogatories are plainly relevant. For example, OpenAI's response to FTC Interrogatory No. 15 will directly address whether OpenAI used Plaintiffs' works to train one or more of its LLMs. Moreover, the burden is minimal. OpenAI need only produce responses it has already shared with the FTC. *See, e.g., Waldman v. Wachovia Corp.*, 2009 WL 86763, at *1-2 (S.D.N.Y. Jan. 12, 2009) (ordering production of materials provided to regulators because the "burden is slight when a defendant has 'already found, reviewed and organized the documents'"); *Sticht v. Wells Fargo Bank, N.A.*, No. 3:20-cv-1550, 2023 WL 2206641, at *5 (D. Conn. Feb. 24, 2023) (burden of producing documents that only require uploading is trivial).

OpenAI's Objections Are Without Merit. Despite this, OpenAI makes a blanket objection that material produced to the government is not discoverable. But OpenAI cannot "withhold otherwise discoverable documents from production in private civil litigation because they were previously submitted to" a federal agency. *Michelo v. Nat'l Collegiate Student Loan Tr.* 2007-2, No. 18-CV-1781, 2020 WL 9423921, at *1 (S.D.N.Y. Aug. 31, 2020). Courts routinely order the production of material provided to the government when it is relevant to a litigation. *See, e.g., id.* Nor is responsive material shielded from discovery merely because the government investigation has a different focus. Discoverable materials "encompass[] investigations of broader practices or issues that are not explicitly tied to [the subject matter of] this case but are nevertheless pertinent." *Ft. Worth Emps.' Ret. Fund v. J.P. Morgan Chase & Co.*, 297 F.R.D. 99, 111 (S.D.N.Y. 2013). Here, Plaintiffs seek only the relevant portions of the materials produced to the FTC. They are entitled to those regardless of the investigation they were produced in.³

OpenAI has suggested that its FTC interrogatory responses should count against Plaintiffs'

¹ OpenAI has represented that it has not produced documents to the FTC. As such, Plaintiffs are not moving to compel OpenAI's production to the FTC but may move separately once OpenAI makes that production.

² Plaintiffs also have served multiple requests seeking OpenAI's organizational structure and employees with knowledge of OpenAI's LLMs. *See e.g.* Ex. C, RFP Nos. 14-15, 30-31.

³ That principle is entirely consistent with OpenAI's only citation, *New Jersey Carpenters Health Fund v. DLJ Mortg. Cap., Inc.*, 2012 WL 13135408 (S.D.N.Y. Mar. 2, 2012). There, the court denied a request for *all* documents supplied to the government in an investigation, because the investigation, and thus the plaintiff's request, was "far broader" than the plaintiff's claims. *Id.* at *1. Here, by contrast, Plaintiffs are seeking production of responses to less than half of the FTC's interrogatories, and have limited their request to only the most relevant ones.

Hon. Sidney H. Stein
May 6, 2024
Page 3

limit under Rule 33, even though those responses were generated as part of an investigation not subject to the Federal Rules of Civil Procedure. This argument is meritless. The presumptive limit in Rule 33 exists “because [interrogatories] are more burdensome than other discovery methods” and plainly does not apply to the trivial burden of *re-producing* responses served in other proceedings. *Spectrum Dynamics Med. Ltd. v. Gen. Elec. Co.*, 18-CV-11386, at *5 (S.D.N.Y. Feb. 25, 2021). OpenAI’s written responses to the FTC are relevant and responsive documents within the meaning of Rule 34 and should be produced.

IDENTITIES OF CRITICAL FORMER EMPLOYEES

To help “train” its LLMs, OpenAI created two data sets known as books1 and books2 that together likely contain more than 100,000 published books. Plaintiffs discuss these data sets in their complaint. *See, e.g.*, Dkt. No. 69, ¶¶ 112-122; ¶ 115 (“Some independent AI researchers suspect that Books2 contains or consists of ebook files downloaded from large pirate book repositories”). OpenAI has made clear that it will dispute whether Plaintiffs’ copyrighted works were a part of these data sets or were otherwise used to train its LLMs. *See generally* Dkt. No. 83 (arguing that requesting admission of these facts is improper).

On March 22, 2024, months into Plaintiffs’ effort to obtain basic information about training, OpenAI revealed that it had destroyed *all* of its copies of books1 and books2, and that the two OpenAI researchers responsible for creating these data sets are no longer employed by OpenAI. Ex. D. That same day, Plaintiffs requested that OpenAI identify these former employees, or else produce documents sufficient to show who these employees were. Plaintiffs have repeatedly followed up since then, to no avail, even though documents sufficient to show this information is obviously responsive to Plaintiffs’ requests for production. *See* Ex. B, RFP Nos. 13-14.

Given that OpenAI destroyed the direct evidence of the content of books1 and books2, these former employees are critically important to this case. OpenAI has supplied no basis for withholding their identities, and its failure to do so is hampering the parties’ ability to negotiate custodians for ESI collections. Plaintiffs ask the Court to direct OpenAI to immediately produce documents sufficient to identify these two key witnesses. *Karsch v. Blink Health Ltd.*, No. 17-CV-3880, 2019 WL 2708125, at *7 (S.D.N.Y. June 20, 2019) (ordering a party to “identify the individuals with personal knowledge concerning” deleted data); *Conservation Law Foundation, Inc. v. Shell Oil Co.*, 2023 WL 5434760, at *19 (D. Conn. Aug. 22, 2023) (finding documents sufficient to show organizational structure, including organizational charts, relevant to plaintiff’s claims and defenses and ordering defendant to produce the same).

The content of OpenAI’s training datasets is of paramount importance to this case and is readily available to OpenAI. Yet OpenAI is chronically resisting this basic discovery and needlessly consuming resources in the process. *See e.g.*, Dkt. No. 78 (seeking to compel OpenAI to admit/deny what is in their training data). Plaintiffs respectfully request production of the FTC materials and the names of the former employees that created books1 and books2.

Hon. Sidney H. Stein
May 6, 2024
Page 4

Respectfully,

LIEFF CABRASER HEIMANN
& BERNSTEINS LLP

SUSMAN GODFREY LLP

COWAN, DEBAETS,
ABRAHAMS & SHEPPARD
LLP

/s/ Rachel Geman
Rachel Geman

/s/ Rohit Nath
Rohit Nath

/s/ Scott J. Sholder
Scott J. Sholder