January 17, 2025

E-Filed

The Honorable Robert M. Illman
United States District Court for the Northern District of California
Eureka-McKinleyville Courthouse
3140 Boeing Avenue
McKinleyville, CA, 95519

Re: *In re OpenAI ChatGPT Litigation*; Master File No. 3:23-cv-3223-AMO

Dear Judge Illman:

Plaintiffs and OpenAI jointly submit this letter brief regarding Plaintiffs' request for an order compelling OpenAI's production of the English Colang Dataset. The parties met and conferred on several occasions, and last met on January 13, 2025, but were unable to reach a resolution.

I. Plaintiffs' Statement

Making no claim of burden or commercial sensitivity, OpenAI already produced two key training datasets containing pirated copies of Plaintiffs' works, known as LibGen 1 and 2. Today, Plaintiffs respectfully ask the Court to compel OpenAI to produce a third training dataset called `cc-en-colang-v2-20220131-metadata` (the "English Colang Dataset"), the [redacted] used to train OpenAI's flagship GPT-4 model. OpenAI's last compromise offer—to produce the subset of documents that are at least 20,000 words long—may sound reasonable, but is another attempt to hamstring Plaintiffs, whose claims entitle them to discover all *infringements* of copyrighted works in the dataset, which can be much shorter than 20,000 words.

Because of the size of the English Colang Dataset—Plaintiffs estimate it contains [redacted] scraped web pages—Plaintiffs have been unable to efficiently or effectively examine the English Colang Dataset via the Inspection Protocol despite numerous good-faith attempts to do so. The Inspection Protocol requires Plaintiffs to access training datasets by connecting to a remote server using a laptop in its lawyers' office via a terminal interface. The remote files are stored in a compressed format. Therefore, any dataset search necessarily requires time and computing resources to decompress and transfer the training-data files. The time to perform each search is extraordinary. For instance, when Plaintiffs used code provided by counsel for OpenAI to search a short sequence of words to find one of Plaintiffs' works in the English Colang Dataset, the search had to be cancelled because it was going to take **over 6 hours to complete**. Plaintiffs have experimented with alternatives, but they still do not make the searches practical.

Even though it is highly inefficient to search the English Colang Dataset via the Inspection Protocol, [redacted] the English Colang Dataset could easily be produced on a single portable hard drive. This would resolve all of Plaintiffs' problems: belying OpenAI's claims, its remote, highly-compressed environment will not be faster than a local, non-compressed, indexed one, which is what Plaintiffs will be able to create upon production. Nevertheless, OpenAI has refused Plaintiffs' repeated requests to do so. Because this dataset is critical to discovery in this case, unfortunately Plaintiffs must now burden the Court with an avoidable dispute and seek to compel production. *See Inspection Protocol*. Dkt. 182, ¶ 7(b).

OpenAI Must Produce the English Colang Dataset Natively

The English Colang Dataset is central to this case. It is "[redacted]," for OpenAI's flagship GPT-4 language model. OPCO_NDCAL_0000940 at 948; *see also* OPCO_NDCAL_0638401. It is responsive to multiple discovery requests from Plaintiffs, including RFP No. 2 ("The Training Data for GPT-4."). It consists of about [redacted] files, each containing [redacted] web pages, for a total of [redacted] web pages. Given its size and the manner in which it was acquired, it is very likely that Plaintiffs' works and certainly many class works are contained therein.

OpenAI's refusal to produce the English Colang Dataset natively has no basis in the Federal Rules of Civil Procedure or the ESI protocols in place.[1] Despite Plaintiffs' efforts to work within the confines of the Inspection Protocol, it has proven to be an unreasonable obstacle to searching the English Colang Dataset. The Inspection Protocol specifies that "[t]he secured computer will be equipped with tools that are <u>sufficient for viewing and searching</u> the Training Data made

[1] Nor does the requested dataset fall under the exclusion in the ESI order for "Data that is inaccessible within the meaning of [Rule] 26(b)(2)(B)." ¶ 9(f).

available for inspection." Dkt. 182, ¶ 7(b) (emphasis added); *see also id.* ("Defendants will reasonably cooperate with Plaintiffs to address any technical concerns Plaintiffs may have regarding the hardware and software that is provided to conduct the Training Data review.").[2]

As noted above, Plaintiffs have encountered significant difficulties with meaningfully interrogating the English Colang Dataset via the Inspection Protocol. These difficulties arise largely because the dataset is stored on a remote server, in a compressed format, and decompression and download of the data via WiFi dramatically increases the overhead for performing any search. If the English Colang Dataset is produced, Plaintiffs will never experience these problems again with this dataset. Moreover, Plaintiffs will be able to build indexes of the English Colang Dataset using custom tools that will allow fast searching. Plaintiffs have done this successfully with their produced copies of the LibGen 1 and LibGen 2 datasets. Plaintiffs have every reason to expect the same success with a produced copy here.

For its part, OpenAI has not denied that Plaintiffs have encountered problems with the Inspection Protocol. Largely, OpenAI has suggested that Plaintiffs' attorneys are just doing it wrong, or that a technical expert should be able to accomplish searches faster than reviewing attorneys. But the Inspection Protocol doesn't require Plaintiffs to engage an expert to conduct basic discovery. And anyway, Plaintiffs *did* engage a technical expert—Dr. Crista Lopes, a computer-science professor at UC Irvine— who encountered the same problems with glacially slow search times. Plaintiffs have tried other approaches too. For example, based on a suggestion by defense counsel to employ more "parallel processing" in their search queries, Plaintiffs did so. In one case, a single search still took almost an hour and had to be cancelled because OpenAI wanted Plaintiffs out by 5:00 pm.[3] In another case, the remote training-data environment completely crashed, and Plaintiffs had to end their inspection hours earlier than expected. In response, OpenAI's counsel wrote "I wish you hadn't left!" Plaintiffs shouldn't have to waste limited inspection days coping with technical obstacles. The purpose of the Inspection Protocol was to facilitate discovery, not inhibit it; and it expressly envisioned that it could be changed if it was. In response to these problems, counsel for OpenAI stated that he "asked ChatGPT to write a short [P]ython script to accomplish the same search that you attempted to run and it was significantly faster," but when Plaintiffs asked to try that script, Counsel refused to provide it. OpenAI has also suggested that Plaintiffs work with OpenAI engineers to construct searches. But such cooperation on scores of specific text searches is not only outside the requirements of the Protocol, it would imperil attorney work-product privilege, unlike custodial ESI searches which are typically general and generic. And given the recent data-deletion issue in OpenAI's ongoing suit against the *New York Times*, Plaintiffs' counsel should not have to rely on such a process.[4]

[2] The Protocol also provides for print requests for portions of the data for review. Plaintiffs requested the production of the English Colang Dataset under this protocol also. Dkt. 182, ¶ 7(h).

[3] Plaintiffs have faced other issues such as OpenAI's failure to respond to an inspection request, blacked out dates, delays in scheduling meet and confers, and demands beyond the Protocol. Defendants mention cancellations, but this only confirms the point: coordinating expert attendance and travel makes this process cumbersome and no longer suited to the needs of the case.

[4] Plaintiffs understandably don't want to take OpenAI's word as to what "other plaintiff groups" have done. "This week, the Times alleged that OpenAI's engineers inadvertently erased data the paper's team spent more than 150 hours extracting as potential evidence." https://www.wired.com/story/new-york-times-openai-erased-potential-lawsuit-evidence.

OpenAI's Production of the English Colang Dataset Is Not Burdensome

Native production will substantially increase efficiency and impose no burden on OpenAI. OpenAI already produced its two datasets in their native format, and Plaintiffs had no issues searching them efficiently. *See Woodard v. Labrada*, 2017 WL 10702139, at *4 (C.D. Cal. Apr. 19, 2017) (finding burden minimal). OpenAI's concerns about sensitivity (about this "publicly" available data)[5] are addressed by the Protective Order, which ensures that any sensitive data will be adequately protected. OpenAI has shipped numerous documents to Plaintiffs under that Order and no issues arose. And anyway, OpenAI does not explain why producing stale training data would implicate any sensitive business needs. After all, OpenAI insists that it trains each successive GPT model on new data. *See infra* (re: LibGen no longer being used). Producing this data thus bears little real risk and can, contrary to OpenAI, be produced on a drive with enterprise grade security. Finally, any burden on OpenAI pales in comparison to the burden on Plaintiffs of requiring additional experts to conduct each and every search of OpenAI's datasets from within the remote inspection room. *See Energy Mgmt. Collab., LLC v. Darwin Tech LLC*, 2024 WL 2335629, at *4–5 (C.D. Cal. Apr. 25, 2024) (compelling production in native form).

Because the Inspection Protocol is not adequate here, Plaintiffs request that OpenAI be compelled to produce the Dataset. *See Rambus Inc. v. Hynix Semiconductor Inc.*, 2007 WL 9653194, at *5–6 (N.D. Cal. Sept. 25, 2007) (compelling native production); *Corker v. Wholesale*, 2020 WL 1987060, at *1–2 (W.D. Wash. Apr. 27, 2020) (production not efficient).

II. DEFENDANTS' STATEMENT

Plaintiffs seek to compel production of [redacted] irrelevant records from a proprietary dataset—[redacted] of data—ignoring the agreed-upon and court-ordered Inspection Protocol and OpenAI's reasonable proposed solutions to their concerns. If Plaintiffs want to search more efficiently, the solution is ***not*** a burdensome export that moves [redacted] records to a less secure and far slower system. Instead, the solution is to work together to efficiently use the inspection environment, which OpenAI has repeatedly offered to Plaintiffs, and which OpenAI has done successfully with others in co-pending cases. OpenAI also offered a generous compromise to Plaintiffs' demands: native production of ***any*** record that could possibly be a book (based on length), which is more than Plaintiffs can reasonably demand under the Federal Rules.

Background. The text datasets that OpenAI uses to train foundational large-language models are staggeringly enormous. To date, OpenAI estimates it has made over [redacted] of data available for inspection under the Inspection Protocol. Here, Plaintiffs focus on a particular dataset that contains text extracted from public websites. Plaintiffs' apparent theory is that the third-party foundation that collected the raw website data may have crawled websites that published their books, and therefore the proprietary dataset that OpenAI built using that data might also contain some small number of books. Just as an email repository contains many discrete mailboxes, and each mailbox contains many discrete messages, training datasets are subdivided into many distinct records. The disputed dataset contains [redacted], totaling [redacted], split across [redacted] to improve performance.

[5] Common Crawl is a "501(c)(3) . . . that crawls the web and *freely provides its archives and datasets to the public*." https://en.wikipedia.org/wiki/Common_Crawl (emphasis added).

In response to early RFPs that sought production of ***all*** training data used for three flagship LLMs, OpenAI objected that producing this dataset, and many others implicated by the RFPs, would need to be done by inspection so that Plaintiffs could identify the subsets of data that are actually relevant and proportional to their claims. Plaintiffs then stipulated to a protocol that allowed them to inspect the data as it is kept in the ordinary course on an OpenAI server, and OpenAI agreed to install additional software or program files that Plaintiffs' needed and to produce relevant records identified by Plaintiffs' searches. Dkt. No. 182 at ¶ 7.

Since training data at this scale cannot be searched without significant computing hardware, OpenAI specially built an enterprise-grade virtual machine with 64 CPUs, 128 gigabytes of RAM, and access to high-speed storage. This inspection environment is not only faster than anything Plaintiffs could build locally, it is also protected by OpenAI's security and firewall systems, which is critical given the sensitivity of OpenAI's datasets. OpenAI's counsel accesses the datasets through the exact same system, and OpenAI's engineers who work on these types of datasets also use similar network-restricted servers.

Plaintiffs' inspection failures. While Plaintiffs now complain that the agreed-upon Inspection Protocol prevents them from "meaningfully interrogating" this dataset, OpenAI has worked much more successfully with other plaintiff groups under inspection protocols nearly identical to the one that Plaintiffs now seek to evade. Indeed, every other plaintiff group has recognized (correctly) that searching such massive, unstructured datasets is materially different from searching typical ESI, which can be loaded and indexed on an e-discovery vendor's platform.

Other plaintiff groups, understanding the challenge of searching this much data, have collaborated with OpenAI to conduct efficient searches. For example, some plaintiffs retained consultants to write scripts to search efficiently (with OpenAI promptly producing relevant hits natively). Other plaintiffs sent OpenAI proposed search strings (based on URLs or expected content) for OpenAI to execute on their behalf. And other plaintiffs have used a hybrid approach, sending some searches to OpenAI and doing some themselves. OpenAI has produced large volumes of native data to all those plaintiffs, which they can further analyze on their systems. If Plaintiffs here had similarly engaged with OpenAI months ago, they would be in the same position. Instead, last October, Plaintiffs briefly engaged a consultant who failed to utilize the massive processing power of the inspection environment, and instead tried to search for books one sentence at a time—the most inefficient method possible. After the consultant failed to appear for a November inspection, and cancelled several more, Plaintiffs turned to non-technical attorneys. These attorney-led efforts were, unsurprisingly, inefficient (as Plaintiffs admit), and in one incident, Plaintiffs' attorney accidentally launched so many processes that OpenAI had to restart a background service (which took a few minutes, though counsel had already walked out). Overall, since Plaintiffs agreed to the Inspection Protocol 116 days ago, they have inspected the datasets just ***six*** times and abandoned or cancelled inspections ***fifteen*** times. (Plaintiffs blame travel issues for their absenteeism, but it was Plaintiffs who demanded that inspections be held "within 25 miles of San Francisco." Dkt. 182 at ¶ 3.)

Plaintiffs also appear confused about how the inspection environment works despite several explanations, but to avoid doubt: Plaintiffs' difficulties have nothing to do with the server being "remote" (in a high-performance datacenter). The relevant network is ***not*** the WiFi in counsel's offices, but the connection between cloud resources, which is hundreds of times faster than WiFi. And the cloud-based storage is much faster than an external hard drive. In sum, the problem is not the hardware or network; it is Plaintiffs' refusal to work cooperatively. For example, when Plaintiffs first theorized that compression was slowing them down, OpenAI decompressed it all to a faster disk. But Plaintiffs did not even bother to return. Moreover, Plaintiffs misunderstand the difficulty of enabling "fast searching." They cite the success that they've had searching two smaller (and not competitively sensitive) datasets: "LibGen1 and LibGen2". But Plaintiffs ignore that the dataset at issue here (`cc-en-colang-v2-20220131-metadata`) has roughly [redacted] more data, and [redacted] more records. That's not apples versus oranges, it's two apples versus a cargo ship full of oranges. The comparison is also inapt because Plaintiffs asserted that ***every*** record in the smaller datasets was relevant and given their small size, the burden of producing them was proportional. In addition, OpenAI stopped using these two smaller datasets ***years ago***, so producing them in full did not raise a significant risk of exposing competitively sensitive data.

OpenAI's collaboration efforts. The Inspection Protocol requires OpenAI to "reasonably cooperate with Plaintiffs to address any technical concerns." And OpenAI has repeatedly offered to, but Plaintiffs have refused to cooperate because it supposedly "would imperil attorney work-product privilege." For example, Plaintiffs refuse to share any search terms that could be used on this dataset, because it would supposedly reveal their mental impressions. But given that Plaintiffs have insisted on demanding their own search terms for other e-discovery searches, their refusal to do so here is inexplicable. *See* N.D. Cal. Guidelines for the Discovery of ESI, Guideline 1.02 (Cooperation) ("The Court expects cooperation on issues relating to the . . . search, review, and production of ESI. The Court notes that an attorney's zealous representation of a client is not compromised by conducting discovery in a cooperative manner."). Plaintiffs also argue that the Inspection Protocol invades their work product, because even if they write their own searches and identify hits for production, OpenAI would get to see those documents when it produces them natively. Nonsense. A producing party always sees the documents it produces, and in any event, Plaintiffs ***agreed*** to this procedure. The only thing that has changed since the parties' stipulation is the appearance of new Plaintiffs' counsel. Finally, Plaintiffs argue that if they find even a single asserted work in a dataset, then OpenAI must produce the entire dataset, no matter how irrelevant the other records may be. But if that were the rule, parties could demand production of entire computers, phones, and mailboxes any time a single responsive document is found on them. That is not how ESI discovery works.

Finally, to avoid burdening the Court with this dispute, OpenAI offered to produce ***any*** record from this dataset if it is at least 20,000 words long (~20-to-25% of a typical book), regardless of whether it matches an asserted work: estimated at [redacted] records. Contrary to Plaintiffs' assertions, producing [redacted] records would be far less burdensome than [redacted] records, would put far less of OpenAI's data at risk of being exposed to competitors, and would also be easier for Plaintiffs to search. Unfortunately, Plaintiffs rejected this offer immediately.[6]

[6] If the Court is inclined to compel production of the `cc-en-colang-v2-metadata` dataset in full, the Court should first order the parties to meet and confer on a protocol for securely transferring and hosting the data. This dataset cannot be exported to a hard drive. OpenAI Security would need to develop an encrypted transfer method, along with minimum security requirements for whatever environment Plaintiffs plan to use (which they have refused to disclose to OpenAI).

By: /s/ *Thomas E. Gorman*
Thomas E. Gorman

Thomas E. Gorman
KEKER, VAN NEST & PETERS LLP
633 Battery Street
San Francisco, CA 94111-1809
Telephone: (415) 391-5400
Facsimile: (415) 397-7188
Email: tgorman@keker.com

Elana Nightingale-Dawson (admitted pro hac vice)
LATHAM & WATKINS LLP
555 Eleventh St., NW, Suite 1000
Washington, DC 20004
Telephone: (202) 637-2200
Email: elana.nightingaledawson@lw.com

John R. Lanham
MORRISON & FOERSTER LLP
12531 High Bluff Dr, Suite 100
San Diego, CA 92130
Telephone: (858) 720-5100
Fax: (858) 720-5125
Email: jlanham@mofo.com

Attorneys for OpenAI, Inc.

By: /s/ *Joshua M. Stein*
Joshua M. Stein

BOIES SCHILLER FLEXNER LLP
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

Maxwell V. Pritt (SBN 253155)
Joshua M. Stein (SBN 298856)
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(415) 293-6800
mpritt@bsfllp.com
jstein@bsfllp.com

Jesse Panuccio (*pro hac vice*)
1401 New York Ave, NW
Washington, DC 20005
(202) 237-2727
jpanuccio@bsfllp.com

Evan M. Ezray (*pro hac vice*)
401 East Las Olas Blvd. Suite 1200
Fort Lauderdale, FL, 33301
(954) 377-4237
eezray@bsfllp.com

JOSEPH SAVERI LAW FIRM, LLP
Joseph R. Saveri (SBN 130064)
jsaveri@saverilawfirm.com
Cadio Zirpoli (SBN 179108)
czirpoli@saverilawfirm.com
Christopher K.L. Young (SBN 318371)
cyoung@saverilawfirm.com
Holden Benon (SBN 325847)
hbenon@saverilawfirm.com
Aaron Cera (SBN 351163)
acera@saverilawfirm.com
601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940

Matthew Butterick (SBN 250953)
mb@buttericklaw.com

	1920 Hillhurst Avenue, #406 Los Angeles, CA 90027 Telephone: (323) 968-2632 Facsimile: (415) 395-9940 **CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP** Bryan L. Clobes (*pro hac vice*) Alexander J. Sweatman (*pro hac vice*) Mohammed A. Rathur (*pro hac vice*) 135 South LaSalle Street, Suite 3210 Chicago, IL 60603 Tel: (312) 782-4880 bclobes@caffertyclobes.com asweatman@caffertyclobes.com mrathur@caffertyclobes.com *Counsel for Individual and Representative Plaintiffs and the Proposed Class*

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(h)

I hereby attest that I obtained concurrence in the filing of this document from each of the other signatories. I declare under penalty of perjury that the foregoing is true and correct.

Dated: January 17, 2025

BOIES SCHILLER FLEXNER LLP

/s/ Joshua M. Stein
Joshua M. Stein

Attorneys for Plaintiffs