Exhibit 18

JOSEPH C. GRATZ (CA SBN 240676)
JGratz@mofo.com
TIFFANY CHEUNG (CA SBN 211497)
TCheung@mofo.com
JOYCE C. LI (CA SBN 323820)
JoyceLi@mofo.com
MELODY E. WONG (SBN 341494)
MelodyWong@mofo.com
MORRISON & FOERSTER LLP
425 Market Street,
San Francisco, California 94105-2482
Telephone: (415) 268-7000
Facsimile: (415) 268-7522

ROSE S. LEE (CA SBN 294658)
RoseLee@mofo.com
ALEXANDRA M. WARD (CA SBN 318042)
AlexandraWard@mofo.com
MORRISON & FOERSTER LLP
707 Wilshire Boulevard
Los Angeles, California 90017-3543
Telephone: (213) 892-5200
Facsimile: (213) 892-5454
[CAPTION CONTINUED ON NEXT PAGE]

Attorneys for Defendants
OPENAI, INC., OPENAI, L.P., OPENAI OPCO, L.L.C., OPENAI GP, L.L.C., OPENAI STARTUP FUND GP I, L.L.C., OPENAI STARTUP FUND I, L.P., AND OPENAI STARTUP FUND MANAGEMENT, LLC

ANDREW M. GASS (SBN 259694)
Andrew.Gass@lw.com
JOSEPH R. WETZEL (SBN 238008)
Joseph.Wetzel@lw.com
LATHAM & WATKINS LLP
505 Montgomery Street, Suite 2000
San Francisco, California 94111
Telephone: (415) 391-0600

ROBERT A. VAN NEST (SBN 84065)
rvannest@keker.com
R. JAMES SLAUGHTER (SBN 192813)
rslaughter@keker.com
PAVEN MALHOTRA (SBN 258429)
pmalhotra@keker.com
MICHELLE YBARRA (SBN 260697)
mybarra@keker.com
NICHOLAS S. GOLDBERG (SBN 273614)
ngoldberg@keker.com
THOMAS E. GORMAN (SBN 279409)
tgorman@keker.com
KATIE LYNN JOYCE (SBN 308263)
kjoyce@keker.com
CHRISTOPHER SUN (SBN 308945)
csun@keker.com
KEKER, VAN NEST & PETERS LLP
633 Battery Street
San Francisco, CA 94111-1809
Telephone: (415) 391-5400
Facsimile: (415) 391-7188

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

IN RE OPENAI CHATGPT LITIGATION

This document relates to:

All Actions

Master File No. 3:23-cv-03223-AMO

OPENAI'S RESPONSES TO PLAINTIFFS' SECOND SET OF INTERROGATORIES (NOS. 17-20)

DESIGNATED HIGHLY CONFIDENTIAL – ATTORNEYS' EYES ONLY

overly broad, unduly burdensome, and not proportional to the needs of the case, including because it includes non-textual material, material only "considered for use," and "material used to achieve or improve its performance during any phase of its development." OpenAI further objects to Plaintiffs' definition of "Training Data" as vague and ambiguous as to the meaning of the terms "pre-training, mid-training, post-training, validation, testing, fine-tuning, and alignment." OpenAI further objects to Plaintiffs' definition of "Training Data" as overbroad because it is not limited to datasets used to train the relevant models used to power ChatGPT. OpenAI objects to this Interrogatory as compound. OpenAI objects to this Interrogatory to the extent that it calls for the disclosure of information protected by the attorney-client privilege, the work-product doctrine, the common interest doctrine, the settlement privilege, and/or any other privilege or immunity from disclosure recognized by law.

Subject to and without waiving the foregoing general and specific objections, OpenAI responds that it is willing to meet and confer regarding the scope of information sought by this Interrogatory.

INTERROGATORY NO. 18:

Describe in detail "Logistic Regression" or "Logistic Regression Filter" referenced in OPCO_NDCAL_0855166.

RESPONSE TO INTERROGATORY NO. 18:

OpenAI objects to this Interrogatory to the extent it seeks documents and information that are not relevant to any claim or defense in this litigation, including to the extent it seeks information about models other than those used to power ChatGPT, i.e., GPT-3.5, GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo. OpenAI objects to this Interrogatory as overly broad and unduly burdensome because it is not sufficiently limited by time. OpenAI objects to this Interrogatory as overly broad, unduly burdensome, and disproportionate to the needs of this case to the extent it asks for a description of "Logistic Regression." OpenAI objects to this Interrogatory as compound. OpenAI objects to this Interrogatory to the extent that it calls for the disclosure of information protected by the attorney-client privilege, the work-product doctrine,

the common interest doctrine, the settlement privilege, and/or any other privilege or immunity from disclosure recognized by law.

Subject to and without waiving the foregoing general objections, OpenAI responds that the "Logistic Regression Filter" identified in OPCO_NDCAL_0855166 (hereinafter, the "LR Filter") refers to, as its name suggests, a filter that OpenAI had used in the past for separating out those portions of content downloaded from the Common Crawl Foundation ("Common Crawl") that are not likely to be usable for LLM training.

The vast majority of data that is collected and distributed by Common Crawl is unusable for training OpenAI's LLMs (though it may have value for other purposes).

- According to Common Crawl, it gathers raw snapshots of web pages stored primarily in WARC files (Web ARChive format). *See Get Started*, Common Crawl, https://commoncrawl.org/get-started ("Common Crawl currently stores the crawl data using the Web ARChive (WARC) Format."); *see also The WARC Format 1.1*, IIPC Community Resources, https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/. These files contain a broad range of content, including raw HTML, JavaScript, CSS, metadata, and multimedia content, which (in this raw format) presents significant issues for training LLMs. *See, e.g.*, https://gist.github.com/pjox/4de9fb8962b715fe73bbe09d70b2e94e?h=1#file-saturn-warc (example of WARC file).
- First, text training data for the purpose of training models of natural language typically needs to be plain text, not HTML or script tags. S. Jain Dep. Tr. at 190:17-191:9. Text from web data containing complex structures, embedded scripts, navigation menus, or irrelevant boilerplate content cannot be easily extracted.
- Second, a huge portion of the HTML content includes templates, navigation text, ads, cookies notifications, site-navigation bars, headers, footers, and duplicated content, which reduces the proportion of useful text available on a page. Penedo et al., *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*, Falcon LLM (Jun 1, 2023), https://arxiv.org/abs/2306.01116 (noting that even a

preprocessed version of Common Crawl, which is supposed to include only plain text, still includes "undesirable navigation menus, ads, and other irrelevant texts").

- Third, beyond formatting challenges, openly crawled web content inherently includes other unusable content like spam or machine-generated pages designed for SEO (Search Engine Optimization), resulting in repetitive, keyword-stuffed, unnatural-sounding text. *See, e.g.,* Rae et al., *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*, DeepMind (Jan. 21, 2022), https://arxiv.org/abs/2112.11446 ("The vast majority of text found on the web is of insufficient quality to be useful for language model training. For example, many web pages contain primarily automatically generated content, or text that is not intended for human consumption (such as keywords for search-engine optimisation)."); *see also* Wenzek et al., *CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data*, (LREC 2020), https://aclanthology.org/2020.lrec-1.494/ (proposing one "pipeline" to extract usable training data from Common Crawl).

To extract usable (referred to as "quality" in OPCO_NDCAL_0855166) text data for LLM training purposes from Common Crawl, OpenAI has employed a complex data processing pipeline. In the past, one small component of this pipeline, which OpenAI has not used in several years, involved the use of the LR Filter. The GPT-3 team developed this method as part of the pipeline that filtered the approximately 45TB of compressed plaintext in Common Crawl down to 570GB. *See* T Brown et al., *Language Models are Few-Shot Learners*, OpenAI (July 22, 2020), https://arxiv.org/abs/2005.14165, at 8 (noting that "unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets").

OpenAI has made the LR Filter source code available for inspection, and to the extent this Interrogatory seeks a description of the operation of the LR Filter on a line by line basis at the source code level, OpenAI identifies pursuant to Federal Rule of Civil Procedure 33(d) the source code and all related history that has been made available for inspection on the secured source code computer (*see, e.g.*, "/home/plaintiff_jslf/codev3/web-text/domain_classifier/", "/home/plaintiff_jslf/codev3/data_tinder_tinder/data/tinder/tinder/scripts/filtered_cc",

"/home/plaintiff_jslf/codev3/data_tinder_tinder/spark/tinder/tinder/filtered_cc", and "/home/plaintiff_jslf/codev3/data_tinder_tinder/spark/tinder/tinder/scripts/filtered_cc"). OpenAI has also produced documents and communications related to the development of the LR Filter. At a high level, the original version of the filter worked as follows:

1. "Good" data was selected, where "good" meant only that the data were believed to contain less of the format and type of content that makes unprocessed Common Crawl data difficult or undesirable to use for LLM training (e.g., unreadable data, repetitive text, machine-generated spam, etc.). Based on source code, this data appears to have been drawn from sources such as webtext, webtext2, Wikipedia, Library Genesis, Amazon reviews, Yelp reviews, arXiv, and Project Gutenberg. *See* OPCO_SRC_CODE00000001. This data was loaded into a data frame.
2. Common Crawl data was loaded into a separate data frame to serve as negative examples.
3. A merged data frame was created, but if the text column for a given row was longer than 10,000 characters, then it was truncated to be no more than 10,000 characters. A comment in the original code notes, "We're throwing out most of the document, so take a random subsample rather than prefix" (that is, don't always sample from the first 10,000 characters). This random subsampling process ensures that the sample was not biased towards content that is more likely to appear at the beginning of documents. The commenter also noted that, "For libgen this helps to eliminate copyright info from start of document," which reflects the fact that when *developing a filter* to help *differentiate* between (a) unreadable, unusable, repetitive, or machine-generated data, and (b) potentially usable language data, the examples of potentially usable language data should not *always* be drawn from repetitive, boilerplate text often found at the beginning of documents (*i.e.,* "front matter"). But, since this was a limited subsampling process, some samples would continue to contain such text. Importantly, nothing is ever removed from the underlying data (or the good_df data frame) by this subsampling process.
4. A Spark Machine Learning pipeline was created.

a. The Spark Tokenizer function was used to split text based on whitespace. For example: [“The”, “Spark”, “Tokenizer”, “function”, “was”, “used”, “to”, “split’, “text”, “based”, “on”, “whitespace”].

b. The Spark HashingTF function was used to create a data structure that does not retain any of the sentences from the original documents, and instead counts the frequency of hashes. Specifically, the function creates a vector of over 1 million features, hashes each word to turn it into a 32-bit integer (e.g. -1,718,199,816 or 2,001,177,191), maps each hash to an index in the vector, and increments the relevant index by 1. For example, [2**21,[0,1000,230191,972919,1653718,1999999],([1.0,3.0,1.0,5.0,1.0,7.0])] indicates that the vector has 2,097,152 features, the non-zero locations are at indices 0, 1000, 230191, etc., and the values at those indices are 1,3,1,5,1,7 respectively. The temporary data structure created by this process cannot be reversed to re-create the original sampled text.

c. The vector output of the HashingTF function was passed to a logistic-regression tool, along with the relevant label (0 for data from raw Common Crawl, 1 for data from the curated datasets).

5. The LR Filter—having been trained on sparse vectors hashed from curated datasets and sparse vectors hashed from Common Crawl data—is then used to score a large corpus of Common Crawl data. The LR Filter outputs higher scoring documents, but a randomizing element is used to include some lower scoring documents as well. After the code has run to completion, a new, filtered version of the Common Crawl data is created.

This LR Filter was just one historical step in a complex pipeline to filter and process data. The document cited in the interrogatory, OPCO_NDCAL_0855166, refers to a revised version of the LR Filter that was developed as one step in the processing and filtering used to create cc-filtered-a3-06fzd-langb-2021-06. The document specifically refers to git source-code commit 412055182efe23ae6232d314c74a82b489fe0698 (which corresponds to commit

Dated: March 17, 2025

MORRISON & FOERSTER LLP

By: */s/ John R. Lanham*
John R. Lanham

JOSEPH C. GRATZ
JGratz@mofo.com
TIFFANY CHEUNG
TCheung@mofo.com
JOYCE C. LI
JoyceLi@mofo.com
MELODY E. WONG
MelodyWong@mofo.com
MORRISON & FOERSTER LLP
425 Market Street
San Francisco, California 94105-2482
Telephone: (415) 268-7000
Facsimile: (415) 268-7522

ROSE S. LEE
RoseLee@mofo.com
ALEXANDRA M. WARD
AlexandraWard@mofo.com
MORRISON & FOERSTER LLP
707 Wilshire Boulevard
Los Angeles, California 90017-3543
Telephone: (213) 892-5200
Facsimile: (213) 892-5454

JOHN R. LANHAM
JLanham@mofo.com
SARAH J. VANDERVALK
SVandervalk@mofo.com
MORRISON & FOERSTER LLP
12531 High Bluff Drive, Suite 100
San Diego, CA 92130
Telephone: (858) 720-5100
Facsimile: (858) 523-2803

CAROLYN M. HOMER
CMHomer@mofo.com
MORRISON & FOERSTER LLP
2100 L Street, NW, Suite 900
Washington, D.C. 20037
Telephone: (202) 887-1500
Facsimile: (202) 887-0763

Attorneys for Defendants
OPENAI, INC., OPENAI, L.P., OPENAI OPCO, L.L.C., OPENAI GP, L.L.C., OPENAI STARTUP FUND GP I, L.L.C., OPENAI STARTUP FUND I, L.P., AND OPENAI STARTUP FUND MANAGEMENT, LLC

LATHAM & WATKINS LLP

By: */s/ Elana Nightingale Dawson*
Elana Nightingale Dawson

ANDREW M. GASS
Andrew.Gass@lw.com
JOSEPH R. WETZEL
Joseph.Wetzel@lw.com
LATHAM & WATKINS LLP
505 Montgomery Street, Suite 2000
San Francisco, California 94111
Telephone: (415) 391-0600

SARANG VIJAY DAMLE (*pro hac vice*)
Sy.Damle@lw.com
MICHAEL A. DAVID (*pro hac vice*)
Michael.David@lw.com
ELANA NIGHTINGALE DAWSON (*pro hac vice*)
Elana.Nightingaledawson@lw.com
LATHAM & WATKINS LLP
555 Eleventh Street, NW, Suite 1000
Washington, D.C. 20004
Telephone: (202) 637-2200

ALLISON L. STILLMAN (*pro hac vice*)
Alli.Stillman@lw.com
RACHEL R. BLITZER
Rachel.blitzer@lw.com (*pro hac vice*)
HERMAN H. YUE (*pro hac vice*)
Herman.yue@lw.com
LATHAM & WATKINS LLP
1271 Avenue of the Americas
New York, NY 10020
Telephone: (212) 751-4864

Attorneys for Defendants
OPENAI, INC., OPENAI, L.P., OPENAI OPCO, L.L.C., OPENAI GP, L.L.C., OPENAI STARTUP FUND GP I, L.L.C., OPENAI STARTUP FUND I, L.P., AND OPENAI STARTUP FUND MANAGEMENT, LLC

KEKER, VAN NEST & PETERS LLP

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

ROBERT A. VAN NEST
rvannest@keker.com
R. JAMES SLAUGHTER
rslaughter@keker.com
PAVEN MALHOTRA
pmalhotra@keker.com
MICHELLE YBARRA
mybarra@keker.com
NICHOLAS S. GOLDBERG
ngoldberg@keker.com
THOMAS E. GORMAN
tgorman@keker.com
KATIE LYNN JOYCE
kjoyce@keker.com
CHRISTOPHER SUN
csun@keker.com
KEKER, VAN NEST & PETERS LLP
633 Battery Street
San Francisco, CA 94111-1809
Telephone: (415) 391-5400
Facsimile: (415) 391-7188

Attorneys for Defendants
OPENAI, INC., OPENAI, L.P., OPENAI OPCO, L.L.C., OPENAI GP, L.L.C., OPENAI STARTUP FUND GP I, L.L.C., OPENAI STARTUP FUND I, L.P., AND OPENAI STARTUP FUND MANAGEMENT, LLC

CERTIFICATE OF SERVICE

I declare that I am employed with the law firm of Morrison & Foerster LLP, whose address is 12531 High Bluff Drive, San Diego, CA, 92130. I am not a party to the within cause, and I am over the age of eighteen years.

I further declare that on March 17, 2025, I served a copy of:

OPENAI'S RESPONSES TO PLAINTIFFS' SECOND SET OF INTERROGATORIES (NOS. 17-20)

X **BY ELECTRONIC SERVICE [Fed. R. Civ. P. 5(b)]** by electronically mailing a true and correct copy through Morrison & Foerster LLP's electronic mail system to the email address(es) set forth below, or as stated on the attached service list per agreement in accordance with Fed. R. Civ. P. 5(b).

JOSEPH SAVERI LAW FIRM, LLP
Joseph R. Saveri
Cadio R. Zirpoli
Christopher K. L. Young
William W. Castillo Guardado
Holden J. Benon
Aaron Cera
Margaux Poueymirou
Melissa Tribble
601 California St., Ste. 1505
San Francisco, CA 94108
jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
wcastillo@saverilawfirm.com
hbenon@saverilawfirm.com
acera@saverilawfirm.com
mpoueymirou@saverilawfirm.com
mtribble@saverilawfirm.com
jslfservice_chatgpt@saverilawfirm.com

CAFFERTY, CLOBES MERIWETHER & SPRENGEL LLP
Bryan L. Clobes
Alexander J. Sweatman
Mohammed Rathur
135 South LaSalle Street, Suite 3210
Chicago, IL 60603
bclobes@caffertyclobes.com
asweatman@caffertyclobes.com
mrathur@caffertyclobes.com

MATTHEW BUTTERICK, ATTORNEY AT LAW
Matthew Butterick
1920 Hillhurst Ave. #406
Los Angeles, CA 90027
mb@buttericklaw.com

VENTURA HERSEY & MULLER, LLP
Daniel J. Muller
1506 Hamilton Avenue
San Jose, CA 95125
dmuller@venturahersey.com

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

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

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

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

Counsel for Individual and Representative Plaintiffs and the Proposed Class

I declare under penalty of perjury under the laws of the State of California that the foregoing is true and correct.

Executed at San Diego, CA, this 17th day of March, 2025.

Sarah Vandervalk