GHAJAR
EXHIBIT 50

COOLEY LLP
BOBBY GHAJAR (198719)
(bghajar@cooley.com)
COLETTE GHAZARIAN (322235)
(cghazarian@cooley.com)
1333 2nd Street, Suite 400
Santa Monica, California 90401
Telephone: (310) 883-6400

COOLEY LLP
MARK WEINSTEIN (193043)
(mweinstein@cooley.com)
KATHLEEN HARTNETT (314267)
(khartnett@cooley.com)
JUDD LAUTER (290945)
(jlauter@cooley.com)
ELIZABETH L. STAMESHKIN (260865)
(lstameshkin@cooley.com)
3175 Hanover Street
Palo Alto, CA 94304-1130
Telephone: (650) 843-5000

CLEARY GOTTLIEB STEEN & HAMILTON LLP
ANGELA L. DUNNING (212047)
(adunning@cgsh.com)
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304
Telephone: (650) 815-4131

[Full Listing on Signature Page]

Counsel for Defendant Meta Platforms, Inc.

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

RICHARD KADREY, et al.,

Individual and Representative Plaintiffs,

v.

META PLATFORMS, INC., a Delaware corporation;

Defendant.

Case No. 3:23-cv-03417-VC

DEFENDANT META PLATFORMS, INC.'S FURTHER SUPPLEMENTAL AND AMENDED RESPONSES AND OBJECTIONS TO PLAINTIFFS' FIRST SET OF INTERROGATORIES

Trial Date: None
Date Action Filed: July 7, 2023

Interrogatory, Meta will conduct a reasonable, proportionate search for non-privileged, relevant, responsive information within its possession, custody, or control.

17. In responding to all Interrogatories, Meta will comply with the requirements of the Federal Rules of Evidence and Federal Rule of Civil Procedure 26.

III. OBJECTIONS AND RESPONSES TO INDIVIDUAL INTERROGATORIES

INTERROGATORY NO. 1:

Describe in detail the data You have used to train or otherwise develop the Meta Language Models, Including, for each:

a. How You obtained the data, e.g., by scraping the data, purchasing it from third parties, or by other means;

b. All sources of Data, including any third parties that provided data sets;

c. To the extent the data was derived from publicly available websites, a list of all such websites and, for each, the percentage of the data corpus that is derived from that website;

d. The categories of content included in the data and the extent to which each category is represented in the data corpus (i.e., as a percentage of data used to train the model);

e. All policies and procedures Related to identifying, assessing, vetting and selecting sources of data for the model.

RESPONSE TO INTERROGATORY NO. 1:

Meta incorporates by reference its objections and definitions above, including to the terms "You" and "Meta Language Models." Meta further notes that the capitalized term "Related" is not defined; Meta construes that term coextensively with "concerning."

As an initial matter, Meta objects to this Interrogatory because it consists of multiple, separate Interrogatories, each which count toward Plaintiffs' limit under the Federal Rules. For example, the question about what data used to train a model is separate from how it was obtained, and further, subparts (d) and (e) are not subsumed within and necessarily related to the primary question, and purport to require a calculation of percentages of data, and separate identification of "policies" and "procedures" for (1) identifying, (2) assessing, (3) vetting, and (4) selecting data. This Interrogatory consists of *at least* three Interrogatories, and depending on how it is interpreted,

many more. In answering the Interrogatory, Meta does not waive this objection.

Meta objects to this Interrogatory because, on its face, it does not exclude legal advice or opinions, which are subject to attorney-client privilege and/or attorney work product doctrine, in particular as to subpart (e). Meta will not produce privileged materials or attorney work product.

Meta objects to this Interrogatory as vague and ambiguous as to the term "data," which is undefined. Meta will construe "data" to mean Training Data (as construed above).

Meta objects to this Interrogatory as vague, ambiguous, and unintelligible as to "percentage of that data corpus that is derived from that website" because "data corpus" is undefined, and Meta is accordingly unable to interpret and respond to subpart (c). Even if "data corpus" were defined, the subject matter of subpart (c) would be overbroad, unduly burdensome, and disproportionate to the needs of the case and seeks information that is not relevant to the parties' claims and defenses. Meta will not respond to subpart (c).

Meta objects to the undefined phrase "categories of content, which is vague, ambiguous, and unintelligible.

Meta objects to this Interrogatory to the extent that it seeks information that is not within Meta's possession, custody, or control.

Subject to and without waiving the foregoing objections, and pursuant to the terms of the Protective Order and the ESI Order, Meta responds as follows: Meta incorporates by reference the identification of datasets used to train Llama 1 that is included in the publicly available paper "LLaMA: Open and Efficient Foundation Language Models." Such datasets were used to train Llama 1. Meta will produce a copy of that paper in its forthcoming production pursuant to Rule 33(d).

Meta will conduct a reasonable search for additional non-privileged information or, in accordance with Rule 33(d), documents in Meta's possession, custody, or control, sufficient to show any other datasets used to train the Meta Language Models (as construed above), as well as policies and procedures for identifying, assessing, vetting, and selecting sources of data for those models.

Discovery is ongoing and Meta will also supplement its response to this Interrogatory to

identify the sources of such datasets and general categories of data within them, to the extent that such information is within Meta's possession, custody, or control.

Discovery is continuing and Meta reserves the right to supplement or amend its response at a later time.

Meta's First Supplemental and Amended Response to Interrogatory No. 1:

Subject to and without waiving the foregoing objections, and pursuant to the terms of the Protective Order, Meta responds as follows.

This response is designated as Highly Confidential – Attorney's Eyes Only under the Protective Order.

Based on its reasonable investigation, Meta identifies the following datasets as containing material used to train the Llama Models (as construed above), including pretraining and/or finetuning, as well the locations from which Meta believes they were obtained:

Dataset name	Llama 1	Llama 2	Llama 3	URL or Other Location
Stack Exchange	Yes	Yes	Yes	https://archive.org/details/stackexchange
books3	Yes	Yes	Yes	https://the-eye.eu/public/AI/pile_preliminary_components/books3.tar.gz
Project Gutenberg	Yes	Yes	Yes	https://www.gutenberg.org
Arxiv	Yes	Yes	Yes	https://www.arxiv.org
Github	Yes	Yes	Yes	https://www.github.com
C4	Yes	Yes	Yes	https://www.tensorflow.org/datasets/catalog/c4
CCNet	Yes	Yes	No	https://github.com/speedinghzl/CCNet/blob/master/LICENSE
CC-stories	Yes	Yes	No	https://github/com/tensorflow/models/tree/archive/research/lm_commonsense#1-download-data-files
The Stack	Yes	Yes	Yes	https://huggingface.co/datasets/bigcode/the-stack
Wikipedia	Yes	Yes	Yes	https://en.wikipedia.org/wiki/Wikipedia:Database_download
[redacted]	No	Yes	Yes	[redacted]
	No	Yes	No	
	No	Yes	Yes	

[redacted]			[redacted]
	No	Yes	Yes
	No	Yes	Yes
	No	Yes	Yes
	No	Yes	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes
	No	No	Yes

[redacted]	No	No	Yes	[redacted]	
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
Libgen	No	No	Yes	https://libgen.is	
[redacted]	No	No	Yes	[redacted]	
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		
	No	No	Yes		

[redacted]	No	No	Yes	[redacted]
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	
	No	No	Yes	

The annotations data used for finetuning Llama 2 are identified in Table 6 of the paper titled "Llama 2: Open Foundation and Fine-Tuned Chat Models." Except for the "Meta (Safety & Helpfulness)" data, which was obtained from Meta's vendors, namely, [redacted] these annotations datasets were sourced from publicly available sources, such as Github and Hugging Face. Meta has also entered agreements with [redacted] to provide annotations.

In addition, for Llama 3.1, Meta used publicly available data sourced from [redacted] to train the model, as well as a variety of synthetic data.

The process for selecting datasets for use in pre-training of the Meta Language Models (as construed above) was informed by what data was available, whether the development team believed that the data would help the model achieve optimal results against industry benchmarks, and PXFN review, *i.e.*, cross-functional review by legal, privacy, and/or policy personnel. Each of the above external datasets was required to undergo PXFN review prior to training of the Meta Language Models (as construed above). Any issues related to intellectual property are regarded as legal in nature. Review and consideration of those issues is therefore the responsibility of Meta's legal team, rather than Meta privacy or policy personnel, and is subject to attorney-client privilege and/or work product doctrine.

From the development team's perspective, decisions around which datasets to use for Llama 1 were influenced by the development of other large language models, in particular DeepMind's Chinchilla and the corresponding paper "Training Compute-Optimal Large Language Models." At the time, researchers regarded DeepMind's Chinchilla as state of the art, and the team developing the first version of Llama was motivated to reproduce Chinchilla's results on industry benchmarks (e.g., MMLU, BoolQ, PIQA, etc.) using their own model architecture. Using the same or similar dataset diversity allowed the team to better compare the effectiveness of the respective models. Llama 2 was largely trained on the same datasets as Llama 1.

With respect to Llama 3, whether a particular dataset was used in the training of the model was driven by a number of considerations, including:

- Dataset size – It is understood that LLMs, such as the Meta Language Models (as construed above), require large volumes of text data in order to achieve high performance across

industry benchmarks. In general, the more data the models train on, the better the performance of the model. That is, there is a rough correlation between the number of unique token strings within a dataset and the performance of the models on downstream tasks, such as the ability to answer questions. Accordingly, larger datasets are preferred to smaller datasets.

- Dataset diversity – Datasets with greater diversity of subject matter, a variety of lengths and human/computer languages, and different styles of writing or conversation help enable the models to be more flexible and adaptable to different contexts.
- Dataset quality – Related to the diversity of the dataset is the extent to which undesirable data (such as repetitive data, factually incorrect data, or harmful or toxic data) can be filtered from the dataset without degrading dataset usefulness.

The data mix that will achieve the best results against benchmarks (e.g., MMLU, GSM8K, BoolQ, PIQA, CommonsenseQA, etc.) is difficult to determine in advance. Accordingly, the Meta Language Model (as construed above) development teams performed small scale experiments prior to full scale pre-training to evaluate optimal data mix proportions. Pursuant to Rule 33(d), Meta also refers Plaintiffs to the paper titled "The Llama 3 Herd of Models," published by Meta on July 23, 2024, for further information.

INTERROGATORY NO. 3:

Describe in detail the RLHF process for each Meta Language Model. Include in Your response:

a. Examples of types of experts who write questions and answers for use in RLHF;

b. Examples of questions and answers;

c. An explanation of the rating system or method of evaluation for the Meta Language Model's responses;

d. A description of the RLHF You actually undertook in order to correct or remediate any Meta Language Model's propensity to emit protected expression from its Training Data.

RESPONSE TO INTERROGATORY NO. 3:

Meta incorporates by reference its objections and definitions above, including to the terms

Dated: December 13, 2024

COOLEY LLP

By: */s/ Judd Lauter*
Bobby Ghajar
Mark Weinstein
Kathleen Hartnett
Phillip Morton
Judd Lauter
Elizabeth L. Stameshkin
Matthew Brigham
Colette Ghazarian
Juan Pablo Gonzalez
Cole A. Poppell

LEX LUMINA PLLC
Mark A. Lemley

CLEARY GOTTLIEB STEEN & HAMILTON LLP
Angela L. Dunning

Attorneys for Defendant
META PLATFORMS, INC.

Full Counsel List

COOLEY LLP
PHILLIP MORTON *(pro hac vice)*
(pmorton@cooley.com)
COLE A. POPPELL *(pro hac vice)*
(cpoppell@cooley.com)
1299 Pennsylvania Avenue, NW, Suite 700
Washington, DC 20004-2400
Telephone: (202) 842-7800

COOLEY LLP
MATTHEW BRIGHAM (191428)
(mbrigham@cooley.com)
JUAN PABLO GONZALEZ (334470)
(jgonzalez@cooley.com)
3175 Hanover Street
Palo Alto, CA 94304-1130
Telephone: (650) 843-5000

LEX LUMINA PLLC
MARK A. LEMLEY (155830)
(mlemley@lex-lumina.com)
745 Fifth Avenue, Suite 500
New York, NY 10151
Telephone: (646) 898-2055

VERIFICATION

I, Michael Clark, declare:

I am an employee of Meta Platforms, Inc. ("Meta"), a corporation organized and existing under the laws of Delaware, which is the Defendant in the above-entitled action, and I have been authorized to make this verification on its behalf.

I have read the following documents:

- Meta's Further Supplemental and Amended Responses and Objections to Plaintiffs' First Set of Interrogatories.
- Meta's Further Supplemental and Amended Responses and Objections to Plaintiffs' Second Set of Interrogatories.
- Meta's First Supplemental Responses and Objections to Plaintiffs' Third Set of Interrogatories

I believe, based on personal knowledge or upon information and belief, that those responses are true and correct.

I declare under penalty of perjury under the laws of the United States that the foregoing is true and correct.

Executed at Denver, Colorado on December 13, 2024.

Michael Clark