1

2

3

4          UNITED STATES DISTRICT COURT

5          NORTHERN DISTRICT OF CALIFORNIA

6

7     RICHARD KADREY, et al.,                      Case No.  23-cv-03417-VC  (TSH)

8              Plaintiffs,

9          v.                                       **PUBLIC VERSION OF DISCOVERY
                                                     ORDER AT ECF NO. 366**
10    META PLATFORMS, INC.,                          Re: Dkt. Nos. 352, 353, 354, 355, 356, 357

11             Defendant.

12

13         The parties have a series of discovery disputes.  ECF Nos. 352-57.  The Court rules as

14    follows.

15    **A.      ECF No. 352 (Plaintiffs' RFAs)**

16         **1.      Timeliness**

17         Plaintiffs' motion to compel concerning Meta's RFA responses is timely.  ECF No. 253

18    ("[T]his order clarifies that the deadline to raise disputes regarding additional discovery remains 7

19    days after the close of discovery, as provided by Local Rule 37-3.").

20         **2.      RFAs 3-7, 17, 20, 23, 34, 43, 45-91, 94 and 96**

21         RFAs 3-7, 17, 20, 23, 34 and 43 ask about "copyrighted books," "copyrighted works," and

22    "copyrighted material," and RFAs 45-91, 94 and 96 ask about Plaintiffs' works.  For RFAs 3-7,

23    17, 20, 34, 34 and 43, Meta admits "text from" copyrighted books or works.  For RFAs 45-89,

24    Meta admits "some text from" Plaintiffs' works, and for RFAs 90, 91, 94 and 96, Meta admits

25    "text from" Plaintiffs' works.  Plaintiffs argue these responses are non-responsive, vague and

26    evasive because Meta does not say whether it used a word, a phrase, a sentence, a paragraph, a

27    chapter, or the whole book.  Meta says that providing an exact answer would require an

28    enormously burdensome word-by-word comparison not only for Plaintiffs' books at issue but also

1    for all other copyrighted works in numerous datasets.

2          The Court agrees with Plaintiffs. Meta's argument falsely takes the view that there are

3    only two possible states of knowledge the company can have: (1) it knows that at least one word

4    in the copyrighted work is in the dataset, or (2) through exhaustive, time-consuming forensic

5    analysis, it has determined that the dataset contains an exact copy of the copyrighted work.

6    Arguing that #2 requires an impractical level of work, Meta therefore answers #1. But the Court

7    sees through this false dichotomy. If it looks like a dataset contains a book, a truthful answer is:

8    "admit as to substantially all of the book." Meta is not allowed to use minor quibbles about

9    whether 99% of the book as opposed to exactly 100% of it was included as a pretext for providing

10   no information at all about how much of the book was included.

11         Meta also claims that after having done a reasonable investigation, it concluded that it did

12   not train any Llama model on the entirety of Plaintiffs' books. But that doesn't justify the

13   hopelessly vague "some text from" or "text from" responses. The investigation that Meta says it

14   did should allow Meta to give a better answer than that, such as "admit as to substantially all of

15   the book," "admit as to about half of the book," or "admit as to some text from the book but deny

16   as to substantially all of the book."

17         The Court **GRANTS** Plaintiffs' motion to compel as to these RFAs.

18         **3.      RFAs 7, 16, 19, 22, 26, 35, 39**

19         RFA 7 asked Meta to "[a]dmit that You did not obtain permission or consent from the

20   relevant copyright owners to use all copyrighted books in the Datasets used to train Llama

21   Models." Meta responded that it "admits that one or more Datasets used to train its Llama Models

22   contained text from published and commercially-available versions of one or more copyrighted

23   books for which it did not obtain permission or consent from the relevant copyright owner(s)."

24   That response is defective because RFA 7 asked about all of the datasets used to train Llama

25   models and all of the copyrighted books in them. Responding that "one or more" of the datasets

26   contains text from "one or more" copyrighted books for which Meta did not obtain permission or

27   consent avoids the thrust of the RFA, which is to ask Meta to admit that it did not do this at all.

28   Further, despite the broad sweep of this RFA, the Court doubts any effort is required to answer it.

RFAs 16, 19 and 22 asked Meta to admit that it used Books3, Library Genesis and The Pile as datasets to train one or more Llama models. In response, Meta admits that it used "a portion" or "some content" from those datasets. The Court agrees with Plaintiffs that these responses are vague and Meta's qualifications are meaningless. Meta must provide some estimate or approximation so that Plaintiffs can know if Meta is almost all admitting, mostly admitting, partly admitting, largely denying, or almost completely denying these RFAs. Right now that's not clear. The Court agrees with Meta that this is an RFA, not a rog, so Meta is not obligated to provide a long narrative response, but its existing responses are inadequate because they do not indicate how much of the RFAs are admitted or denied.

RFA 26 asked about "copyright owners," not "persons," and about "negotiat[ing] licensing of their copyrighted material," not "agreement[s] for access to and use of certain data that may include copyrighted material as training material." Meta must answer the RFA as written.

RFA 35 asked Meta to "[a]dmit that Meta has not provided to Plaintiffs a list of works used in the Datasets used to train Llama Models." Meta denied the RFA as to the Books3 dataset, saying it did produce a list of that content to Plaintiffs. For other datasets, Meta said it is not aware of or in possession of any lists of their content. The Court agrees with Plaintiffs that if Meta doesn't have any lists of works used in other datasets, then obviously it hasn't provided Plaintiffs with them, and Meta should admit that.

RFA 39 asked Meta to "[a]dmit that You have not deleted all copyrighted material in Your possession after it is used for training Llama Models." Meta answers in terms of "training data," and then volunteers that it didn't do so, in part, because of its preservation obligations. But Plaintiffs asked about "copyrighted material," not training data, and they didn't ask for a self-serving partial list of reasons. Meta must answer the RFA as it was asked.

The Court **GRANTS** Plaintiffs' motion to compel as to these RFAs.

**4.      RFAs 38, 44, 98**

These RFAs asked Meta to "[a]dmit that You store copyrighted material for training Llama Models," "[a]dmit that if copyright holders or other content creators demanded that You not use their content to train Your LLM models, then You would not use their content to train Your LLM

3

1     models," and "[a]dmit that you used books sourced from Books3 to train one or more of your

2     large language models." Meta claims not to understand these RFAs, but they are simple and clear,

3     and the Court **ORDERS** Meta to answer them.

4             **5.**      **Other Qualifications**

5           Plaintiffs object that in dozens of the RFA responses, Meta begins its answer by saying

6     "subject to and without waiving the foregoing objections," and concludes with "[e]xcept as

7     expressly admitted, Meta denies the Request." However, the Court concludes that these phrases

8     do not interject uncertainty into Meta's answers and therefore **DENIES** Plaintiffs' request to strike

9     them.

10     **B.**      **ECF No. 353 (Meta's Privilege Logs)**

11          Plaintiffs challenge several hundred entries on Meta's privilege logs. However, the Court

12     agrees with Meta that Plaintiffs have failed to provide a factual basis for *in camera* review of these

13     documents, let alone to simply overrule Meta's claims of privilege. Further, in response to

14     previous joint discovery letter briefs, the Court engaged a time-consuming *in camera* review of

15     Meta's documents, largely concurred in Meta's privilege claims (ECF No. 351, addressing ECF

16     Nos. 309, 334, 336), and found "no indication that Meta is abusing the privilege." ECF No. 351 at

17     7. The Court therefore **DENIES** Plaintiffs' motion.

18     **C.**      **ECF No. 354 (Plaintiffs' Motion to Reopen Depositions)**

19          Plaintiffs move to reopen the depositions of Melanie Kambadur, Sergey Edunov, Joelle

20     Pineau, Eleonora Presani and Nikolay Bashlykov for an additional two hours each, as well as the

21     deposition of Mark Zuckerberg for one additional hour. Plaintiffs state that at 10:00 p.m. on

22     December 13, 2024, the last day of fact discovery, Meta produced 2,404 documents totaling

23     around 21 GB of data. Plaintiffs state that the file paths of these documents indicate that hundreds

24     of them were collected by Meta in June 2024 and hundreds more were collected in September

25     2024. Plaintiffs complain that Meta sat on this pile of "hot" documents right until the end of fact

26     discovery, preventing Plaintiffs from using them in depositions.

27          Meta disagrees. It says that Plaintiffs' first sets of RFPs, which were served in April, were

28     relatively narrow. Meta says it responded to those RFPs in advance of the July substantial

completion deadline. Meta argues that it served responses to Plaintiffs' fourth, fifth and sixth sets of RFPs on September 30, November 8 and November 18. Further, on October 4, 2024, the Court ordered Meta to add five additional ESI custodians. ECF No. 212. Meta contends that its December 13 document production reflects those custodial searches and Meta's updated productions in response to Plaintiffs' fourth through sixth sets of RFPs. Meta does not deny the June and September collection dates, but says it produced documents when they became responsive, which changed in response to the more recent RFPs.

It seems to the Court that if Plaintiffs asked for these documents early in discovery, but Meta withheld them until the last day of fact discovery, then Plaintiffs have a good argument to reopen these depositions. On the other hand, if Plaintiffs waited until very late in fact discovery to ask for these documents, then Plaintiffs do not have a good argument to reopen these depositions.

Well, which is it? Interestingly, Plaintiffs don't say. They do not make the argument that they asked for these documents early in discovery. Meta adamantly argues that these documents were responsive only to the later served RFPs. In response, Plaintiffs do not deny that, but confusingly say that "if Meta withheld these documents because they 'weren't responsive,' that is the worst possible explanation. It means a Meta attorney reviewed these documents long ago (which plainly hit on Meta's original search terms, as shown in Dkt. No. 321-2), but conveniently decided that all the outright admissions about how 'LibGen is a pirated dataset' were not responsive to any of Plaintiffs' 50 original RFPs, including, but not limited to, RFPs about 'The Training Data' for Llamas 1-3 and RFPs about any efforts to license training data from copyright holders." (emphasis omitted).

But the Court does not see why that is a bad explanation, let alone "the worst possible explanation." RFPs for the training data asked for the training data, and RFPs about efforts to license training data asked for that. While the Court understands why Plaintiffs see the newly produced documents as significant, the Court does not see that these documents were responsive to the original RFPs that Plaintiffs refer to. In sum Plaintiffs have not demonstrated that Meta inappropriately delayed producing these documents in discovery. Their motion to reopen the depositions is therefore **DENIED**.

5

**D.      ECF No. 355 (Crime-Fraud)**

Plaintiffs seek *in camera* review of a number of documents, citing the crime-fraud doctrine.  Plaintiffs argue that these documents show that Meta torrented copyrighted works in pirated datasets (violating the CDAFA), stripped works of CMI to conceal their copyrighted status (violating the DMCA), and used them to train Llama (violating the Copyright Act).  Plaintiffs argue that Meta engaged its in-house counsel to approve, conceal, and justify this illegal scheme.

Plaintiffs' crime-fraud argument based on the Copyright Act appears to overlap with the merits of their copyright infringement claim.  Plaintiffs' crime-fraud argument based on the CDAFA and the DMCA overlaps with Plaintiffs' motion for leave to file a third amended complaint, which would add those two claims to the case.  The Court has a serious concern about a discovery motion that basically asks the undersigned magistrate judge to decide this lawsuit (including proposed additions to this lawsuit) on the merits, without a trial, in favor of the Plaintiffs.  That seems to get things out of order.  In a case like this where liability is hotly disputed, the Court is not willing to embrace a crime-fraud theory that requires the Court to decide the contested merits of the case in order to rule on a discovery motion.  Plaintiffs' request is **DENIED**.

**E.      ECF No. 356 (Meta Documents and Data)**

**1.      Torrenting Data**

Plaintiffs move to compel Meta to produce its BitTorrent client, application logs, and peer lists – data that reflects how much Meta downloaded and from where, and how much Meta seeded to the internet.  Plaintiffs says these items are responsive to RFPs 85 and 119.  The Court disagrees.  RFP 85 asked for documents about the *decision* to use Torrent Systems, and that's not what Plaintiffs are seeking.  RFP 119 asked for "[a]ll Documents and Communications, including source code, relating to the processing of copyrighted material used in training Llama Models, including storage and deletion of copyrighted material," and Plaintiffs are not asking for that either.  ECF No. 351 at 2 ("The Court does not see how torrenting is responsive to this RFP, which is about the processing of data, not its acquisition.").  Plaintiffs make a technical argument that torrenting does involve data processing, specifically, a torrented file is downloaded in pieces,

which the client then reassembles into the whole file.  *See ME2 Productions, Inc. v. Bayu*, 2017 WL 5165487, *2 (D. Nev. Nov. 7, 2017) ("Once a peer has downloaded the entire file by downloading each of its pieces, the client reassembles the pieces and the peer is able to view the file as a whole.").  However, that is still just a method of obtaining material, whereas RFP 119 is directed to what Meta does with it.  Plaintiffs' motion to compel is **DENIED**.

### 2.    Supervised Fine-Tuning Data

The Court will address this issue in a separate order in connection with ECF No. 361.

### 3.    Llama 4 and 5 Training Datasets

Plaintiffs move to compel production of the datasets used to train Llama 4 and 5.  Meta says that Plaintiffs did not ask for this in discovery.  Meta observes that Plaintiffs' first three RFPs were for "The Training Data for Llama [1/2/3]," and Meta says there was no similar request for Llamas 4 or 5.

Plaintiffs do not have a good response.  They argue that this data is responsive to RFP 81, but that asked for documents about "the decision" to use shadow datasets for training Llama models, not for data sets themselves.  They also point to RFPs 6-12, which asked for documents and communications with various organizations, not for datasets.  It also appears that a motion to compel on RFPs 6-12 was required to be brought by November 8, 2024.  ECF No. 258.

Accordingly, Plaintiffs' motion to compel is **DENIED**.

## F.    ECF No. 357 (Meta's Rog Responses)

On December 13, 2024, Meta served supplemental responses to Plaintiffs' first, second and third sets of interrogatories.  Plaintiffs move to compel on several of these rogs.

A threshold issue is whether Plaintiffs' challenges to Meta's responses to the first and second sets of rogs are time-barred.  Meta served its responses to the first set of rogs on February 23, 2024 and served its responses to the second set of rogs on September 30, 2024.  Those responses therefore constituted "existing written discovery," and November 8, 2024 was the deadline to raise any disputes about them.  ECF No. 258.  However, as Meta served amended responses on December 13, 2024, the Court finds that those amended responses were not "existing written discovery," and challenges to those responses are timely.  The Court is not going to

compare the December 13 responses to the previous responses to see what changed and only deem

challenges timely if they are to the changes.  An amended response is a new response.  As a

practical matter, Meta was free to revise its responses as much or as little as it chose, and Plaintiffs

could not know the extent of the revisions until they received them.  Accordingly, all of Plaintiffs'

challenges to the December 13 amended responses are timely.

### 1.      Set One

In the first set of rogs, Meta's December 13 supplemental responses included responses to

rogs 1, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14 and 15.  Rogs 1, 3, 4, 5, 7, 8, 10, 12, 13 and 15 asked

about "Meta Language Model" or "Meta Language Models."  Plaintiffs complain that Meta

limited its responses to Llamas 1, 2 and 3 and excluded Llamas 4 and 5.  However, as Plaintiffs

acknowledge, "the ROGS initially defined only Llamas 1-3."  ECF No. 357 at 2.  While the Court

did find that Llamas 4 and 5 are relevant, it did not rewrite Plaintiffs' discovery requests.

Plaintiffs argue that Meta improperly limited the term "agreements" to written contracts

and "training data" to Books3.  Meta responds that it did not stand on either objection in

answering the rogs.  With respect to rog 1, the Court can see that Meta did not stand on its

objection to "training data."  But for rog 5, it sure looks like Meta did stand on the objection to

"agreements."  Accordingly, the Court **GRANTS** Plaintiffs' motion to compel as to rog 5 and

**ORDERS** Meta to answer it by also including oral contracts, arrangements or understandings,

whether formal or informal.  *See* ECF No. 315 at 8.

Rog 1.  Plaintiffs challenge Meta's failure to include Llamas 4 and 5.  However, the Court

has addressed that issue above.

Rog 2.  Meta's December 13 amended responses did not include rog 2.  Plaintiffs seek

confirmation that Meta's August 22, 2024 response continues to be the final response.  Meta so

confirms.

Rogs 3 and 7.  Plaintiffs again argue that Meta's responses do not include Llamas 4 and 5.

However, the Court has addressed that issue above.  Plaintiffs also make a two-sentence argument

that these rog responses do not include sufficient information about fine-tuning.  However, both

responses include Rule 33(d) references to documents and testimony, which Plaintiffs do not

discuss or address.  The Court therefore finds that this motion to compel is not sufficiently

explained.  Moving across the board on three sets of rogs in a single joint discovery letter brief

was not a great idea because it has resulted in many of Plaintiffs' arguments being perfunctory.

Rog 4.  Meta argues that this rog is largely irrelevant because most of the risks at issue

have nothing to do with copyright infringement.  Plaintiffs do not meaningfully respond.

Rog 8.  Plaintiffs argue that Meta's response does not include Llamas 4 and 5.  The Court

has addressed that issue above.

Rog 10.  Plaintiffs argue that Meta limited its definition of "training data" to Books3, but

the Court can see from the cross-reference to rog 1 that this is untrue.  Plaintiffs do not explain the

relevance of the additional information they seek in this rog.

Rog 13.  Plaintiffs again argue that Meta did not include Llamas 4 and 5 in its answer, but

the Court has addressed that issue.  Plaintiffs also argue that Meta failed to list any forbidden

datasets and why they were prohibited.  Given that most of the risks Meta identified in the Llama

2 and 3 papers had nothing to do with intellectual property, the Court finds that Plaintiffs have not

adequately explained the relevance and proportionality of that information.

### 2.      Set Two

Meta's December 13 responses to set two responded to rogs 16 and 17.  Rog 16 asked

Meta to "[s]tate all facts on which you base Your contention that Your conduct constitutes fair use

(17 U.S.C. § 107)."  Rog 17 asked:  "If You or any of Your employees and/or agents intend to

assert the advice of counsel defense, state any and all facts upon which You or any of your

employees and/or agents intend to rely on for that contention."

Plaintiffs state that Meta limits its responses to Llamas 1-3.  The Court has addressed that

issue.

Plaintiffs state that Meta improperly limited the definition of "agreements" to written

contracts.  However, the Court does not see what that has to do with rogs 16 or 17, which do not

ask about agreements.

Plaintiffs also state that Meta improperly limited "training data" to Books3.  However, that

does not appear relevant to rogs 16 and 17, which do not ask about training data.

**3.     Set Three**

Meta's December 13 responses to set three included responses to rogs 19-25.

Plaintiffs state that Meta continues to limit its responses to Llamas 1-3, which here also resulted in not disclosing all datasets it uses with Llamas 4 and 5.  The Court has addressed that issue already.

Rog 19.  Meta's response appears to comply with the parties' agreement.  ECF No. 357-5.

Rog 22 asked:  "Describe any efforts You have made to obtain licenses or any similar permissions to use Shadow Datasets, or the works contained therein, to train Llama Models."  Meta responded:  "Meta has not sought licenses or other consent, and maintains that it did not need licenses or other consent, to use the Third Party Datasets to train the Llama Models, because such use was fair use."  Plaintiffs object to the extraneous information in the response.  However, unlike a narrow, focused RFA, this rog ("Describe any efforts") plainly asked for a narrative response, and it is not out of bounds for a narrative response to say why the litigant did what it did.

Rog 23 asked Meta to "[i]dentify all sources from which You have obtained Shadow Datasets."  Plaintiff contend that Meta's response fails to identify what datasets it obtained from Anna's Archive.  However, that information is not responsive to this rog.

**4.     Conclusion**

The Court **GRANTS** Plaintiffs' motion to compel as to rog 5, as explained above, and otherwise **DENIES** the motion.

**IT IS SO ORDERED.**

Dated: January 2, 2025

THOMAS S. HIXSON
United States Magistrate Judge