Chicken Soup for the Soul, LLC v. Meta Platforms Inc. Doc. 131

United States District Court
Northern District of California

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

CHICKEN SOUP FOR THE SOUL, LLC, Plaintiff, v. META PLATFORMS INC., Defendant.	**DISCOVERY ORDER** Case No. 26-cv-02333-VC (TSH) Re: Dkt. Nos. 125, 126
JOHN CARREYROU, et al., Plaintiffs, v. META PLATFORMS, INC., Defendant.	Case No. 26-cv-03725-VC (TSH) Re: Dkt. Nos. 251, 252
COGNELLA, INC., Plaintiff, v. META PLATFORMS, INC., Defendant.	Case No. 26-cv-04053-VC (TSH) Re: Dkt. Nos. 101, 102

Plaintiffs move to compel concerning their subpoenas to OpenAI and Anthropic. The Court addresses the motions below.

A. OpenAI

Plaintiffs move to compel concerning three subjects: market substitution evidence, licensing evidence, and torrenting evidence.

1. Market Substitution Evidence

On this subject, Plaintiffs move to compel on the following document requests:

Dockets.Justia.com

> 1. All Documents and Communications Concerning the ability or capability of any of Your artificial-intelligence models, including any version of ChatGPT, to generate, continue, summarize, transform, imitate, or reproduce literary content of different types or forms, including fiction (including novels and genre fiction), nonfiction (including biographies, memoirs, essays, and journalism), short-form works (including short stories and articles), textbooks, academic or educational works, reference works, poetry, drama, and instructional or technical works.
>
> 2. All Documents and Communications Concerning any testing, evaluation, benchmark, study, analysis, report, or comparison of the abilities or capabilities identified in Request No. 1, including assessments of output length, quality, coherence, creativity, factual accuracy, genre or style fidelity, similarity to existing works or authors, and the extent to which model-generated output may substitute for or compete with human-authored works.

Plaintiffs also move to compel Rule 30(b)(6) testimony on the following topics:

> 1. The ability or capability of any of Your artificial-intelligence models, including any version of ChatGPT, to generate, continue, summarize, transform, imitate, or reproduce literary content of different types or forms, including fiction (including novels and genre fiction), nonfiction (including biographies, memoirs, essays, and journalism), short-form works (including short stories and articles), textbooks, academic or educational works, reference works, poetry, drama, and instructional or technical works.
>
> 2. Any testing, evaluation, benchmark, study, analysis, report, or comparison of the abilities or capabilities identified in Topic No. 1, including assessments of output length, quality, coherence, creativity, factual accuracy, genre or style fidelity, similarity to existing works or authors, and the extent to which model-generated output may substitute for or compete with human-authored works.

Plaintiffs argue that this evidence is relevant to the market harm element of Meta's fair use defense. The Ninth Circuit has explained that "this factor encompasses both (1) the extent of market harm caused by the particular actions of the alleged infringer, and (2) whether unrestricted and widespread conduct of the sort engaged in by the defendant would result in a substantially adverse impact on the potential market for the original and the market for derivative works." *McGucken v. Pub Ocean Ltd.*, 42 F.4th 1149, 1163 (9th Cir. 2022). Plaintiffs emphasize the language about "conduct of the sort engaged in by the defendant" to show that this factor relates to a category of conduct, not a single defendant.

Factually, OpenAI contends that the use of "shadow libraries" in the training of its models was very limited and took place several years ago. OpenAI asserts that its products have had no

connection to such libraries for years. Thus, OpenAI contends it has not committed "conduct of the sort engaged in by" Meta. Plaintiffs respond by saying they will limit these requests and topics to the models OpenAI concedes were trained on shadow-library materials and their successors.

The relevance of the requested information is unclear. The ability of an AI model to generate literary content – the subject of these requests – would seem to depend on the design of the model. Even if OpenAI and Meta had trained their models on the exact same data sets, their models are still different products made by different companies. The unstated premise of Plaintiffs' motion to compel seems to be that the ability of an AI model to generate literary content is principally attributable to the data sets on which it was trained, rather than to the many other decisions that went into product development. You need that premise before you can conclude that the ability of the OpenAI models that were trained on shadow libraries to generate literary content is relevant to showing whether the widespread use of shadow libraries to train AI would result in a substantial adverse impact on the potential market. Yet that premise is not only unstated and unexplained, it seems shaky because it wishes away every difference in product development between Meta's and OpenAI's models. If, for example, OpenAI's models are better at generating literary content because they are simply a better product, then their ability to generate literary content doesn't show anything relevant to Meta's fair use defense.

Generative AI is a new industry, and different AI products by different companies likely have different strengths and weaknesses. It might be true that in mature product markets such as dishwashers and microwave ovens, the industry has figured out that there is one best way to do things, and all the products on the market today are pretty much the same thing. But generative AI is a new thing, and it seems odd to assume that ChatGPT's ability to perform a specific function (here, generating literary content) has nothing to do with features that may be unique to ChatGPT or to product development decisions that were specific to ChatGPT. To draw a straight line from shadow libraries to the ability to generate literary content, you have to assume that Meta's models and OpenAI's models have no relevant differences. That doesn't seem like a good assumption

The Court understands that Plaintiffs are outsiders looking in and cannot reasonably be

expected to have proof about *why* OpenAI's models are or are not good at generating literary content. But they are still the moving party on this motion to compel, which means they have the burden of persuasion. If the ability or inability of OpenAI's models to generate literary content is not primarily caused by the use of shadow libraries to train AI, but is instead driven by other aspects of product development, then all of the market substitution evidence Plaintiffs are seeking is irrelevant to Meta's fair use defense (and Plaintiffs rely on no other theory of relevance). And Plaintiffs are seeking a lot of information, even if they limit their requests to the models OpenAI concedes were trained on shadow-library materials and their successors. The document requests are a boundless inquiry into literary generation, and the deposition topics cannot be responded to without conducting the document searches. In the absence of a persuasive showing of relevance, the Court concludes that this burdensome discovery is not proportional to the needs of the case. Accordingly, this portion of Plaintiffs' motion to compel is **DENIED**.

2. Licensing Evidence

Here, Plaintiffs move to compel concerning the following document request: "3. All Documents and Communications Concerning Your licensing and efforts to license books and other content for AI training." They also move to compel for Rule 30(b)(6) testimony on the topic: "3. Your efforts to license books and other content for AI training."

Plaintiffs again tie these requests to Meta's fair use defense. Meta has contended that there was no licensing market for Plaintiffs' literary works as training data for AI models, there is no such market today, any particular book has no particular value as training data and is interchangeable with countless other books, and it would be impractical or impossible for AI companies to attempt to negotiate licenses with each individual book rights holder. Plaintiffs say that request 3 and topic 3 are relevant to those issues. Plaintiffs say that beyond what OpenAI has agreed to produce, they will limit the request to unconsummated book-licensing negotiations and to custodians OpenAI identified in the MDL.

For its part, OpenAI has agreed to provide Plaintiffs with a set of executed data access agreements. Beyond that, OpenAI says these requests are unduly burdensome, in part because they request new searches for material that exceed the scope and time period at issue in the MDL.

Here, Plaintiffs run into a legal problem. Judge Chhabria has already ruled that the alleged licensing market that these requests relate to, and whether it exists, is legally irrelevant:

> The plaintiffs' primary theory of market harm is that Meta's unauthorized use of their books for LLM training harms the market for licensing their books for that purpose. The plaintiffs devote nearly all of their discussion of the fourth factor to this theory. The parties therefore go back and forth at length about whether a market for licensing general trade books exists or is likely to develop.
>
> But whether such a market exists or is likely to develop is irrelevant, because this market is not one that the plaintiffs are legally entitled to monopolize. In every fair use case, the plaintiff suffers a loss of a potential market if that potential market is defined as the theoretical market for licensing the use at issue in the case. Therefore, to prevent the fourth factor analysis from becoming circular and favoring the rightsholder in every case, harm from the loss of fees paid to license a work for a transformative purpose is not cognizable.

Kadrey v. Meta Platforms, 788 F. Supp. 3d 1026, 1052 (N.D. Cal. 2025) (cleaned up); *see also Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1032 (N.D. Cal. 2025) ("Authors next contend that training LLMs displaced (or will) an emerging market for licensing their works for the narrow purpose of training LLMs . . . . Even so, such a market for that use is not one the Copyright Act entitles Authors to exploit.").

Plaintiffs contend that Judge Chhabria did not regard licensing as training data for LLMs as wholly irrelevant because in section VI.E of his opinion in *Kadrey*, he observed that "where copying for LLM training isn't fair use, LLM developers (including Meta) won't need to stop using copyrighted works to train their models. They will need only to pay rightsholders for licenses for that training." 788 F. Supp. 3d at 1059. However, in that section of his opinion, he was responding to Meta's argument "that the public interest would be badly disserved by preventing Meta (and other AI developers) from using copyrighted text as training data without paying to do so. Meta seems to imply that such a ruling would stop the development of LLMs and other generative AI technologies in its tracks. This is nonsense." *Id*. (cleaned up). That analysis about the consequences of a fair use determination does not alter the discussion earlier in the opinion that a potential licensing market for AI training is irrelevant to making the fair use determination in the first place. Accordingly, the Court **DENIES** this portion of Plaintiffs' motion to compel.

3. Torrenting Evidence

On this subject Plaintiffs move to compel on the following document requests:

> 4. Documents and Communications sufficient to identify each torrenting application, torrent software version, implementation, library, and/or Torrent Client You used to obtain Data from Online Databases from January 1, 2019 to the present.
>
> 5. All Documents and Communications, including configuration files, settings registries, and similar records, Concerning Your efforts to prevent, limit, or mitigate uploading, reuploading, seeding, peer exchange, leeching, or otherwise making Data available to other users during or after the Torrenting process, including by altering configurations and settings of Torrent Clients, devices, servers, virtual machines, networks, cloud environments, or storage systems.
>
> 6. All expert reports, declarations, affidavits, analyses, rebuttal reports, supplemental and/or reply reports, demonstratives, testimony, exhibits, appendices, schedules, attachments, and other expert material prepared by, for, reviewed by, considered by, relied upon by, submitted by, or attributed to Barbara Frederiksen-Cross in *In Re: OpenAI, Inc. Copyright Infringement Litigation*, 1:25-md-03143, (S.D.N.Y.).

They also move to compel Rule 30(b)(6) testimony on the following topics:

> 4. The identity, versions, configuration, and settings of each Torrent Client You used to obtain Data from Online Databases from January 1, 2019 to the present.
>
> 5. Your efforts to prevent, limit, or mitigate uploading, re-uploading, seeding, peer exchange, or otherwise making Data available to other users during or after the Torrenting process.

Plaintiffs again point to Meta's fair use defense. The observe that Meta claims that using BitTorrent was an efficient and reliable means to obtain datasets, that the datasets were only available in bulk through torrent downloads, and that the potential for a peer to upload data to other peers in the network is an inherent characteristic of the BitTorrent protocol.

OpenAI disputes the relevance of this information. It observes that the technical capabilities of torrent clients can be confirmed by examining those clients. OpenAI also observes there is no evidence that it used the same torrent clients or built comparable corpora to Meta.

The Court thinks OpenAI has the better of the argument. To the extent Meta's fair use defense hinges on the assertion that its use of BitTorrent was the only way BitTorrent can be used, that assertion can be tested by examining the BitTorrent client itself. In this case, singling out

OpenAI to see how it used torrent clients, to serve as evidence of how torrent clients can be used, has a sort of random quality to it. Any user of a torrent client would be relevant in that sense. Why can't Plaintiffs' expert use the torrent clients to show how torrent clients can be used?

Further, to the extent that Meta's fair use defense hinges on the assertion that the particular shadow libraries it downloaded were only available in bulk through torrent downloads, that should be evident from the libraries themselves. More generally, even though OpenAI acknowledges that some of its older models made use of shadow libraries, Plaintiffs have not shown that the corpora OpenAI created were similar to the ones Meta created. Thus, Plaintiffs have not shown that how OpenAI made use of torrenting sheds light on other ways that Meta could have built its corpora.

Finally, even if there were some relevance to these requests, they are still disproportional to the needs of the case. Requests four and five require a vast technical investigation. The Court does not see the utility of rummaging around in the MDL for the items requested in 6. And the deposition topics could only be competently handled after the sweeping technical investigation called for by document requests four and five.

For these reasons, this part of Plaintiffs' motion to compel is **DENIED**.

B. Anthropic

The discovery dispute between Plaintiffs and Anthropic is very similar to the one discussed above concerning OpenAI.

1. Market Substitution Evidence

Plaintiffs move to compel on the following document requests:

> 1. All Documents and Communications Concerning the ability or capability of any of Your artificial-intelligence models, including any version of Claude, to generate, continue, summarize, transform, imitate, or reproduce literary content of different types or forms, including fiction (including novels and genre fiction), nonfiction (including biographies, memoirs, essays, and journalism), short-form works (including short stories and articles), textbooks, academic or educational works, reference works, poetry, drama, and instructional or technical works.
>
> 2. All Documents and Communications Concerning any testing, evaluation, benchmark, study, analysis, report, or comparison of the abilities or capabilities identified in Request No. 1, including assessments of output length, quality, coherence, creativity, factual accuracy, genre or style fidelity, similarity to existing works or

> authors, and the extent to which model-generated output may substitute for or compete with human-authored works.

They also move to compel on the following Rule 30(b)(6) topics:

> 1. The ability or capability of any of Your artificial-intelligence models, including any version of Claude, to generate, continue, summarize, transform, imitate, or reproduce literary content of different types or forms, including fiction (including novels and genre fiction), nonfiction (including biographies, memoirs, essays, and journalism), short-form works (including short stories and articles), textbooks, academic or educational works, reference works, poetry, drama, and instructional or technical works.
>
> 2. Any testing, evaluation, benchmark, study, analysis, report, or comparison of the abilities or capabilities identified in Request No. 1, including assessments of output length, quality, coherence, creativity, factual accuracy, genre or style fidelity, similarity to existing works or authors, and the extent to which model-generated output may substitute for or compete with human-authored works.

Plaintiffs' arguments with respect to market substitution are similar to the ones described above concerning OpenAI. Factually, Plaintiffs observe that *Bartz* made clear that Anthropic trained at least some of its models on shadow libraries. 787 F. Supp. 3d at 1015 ("As Anthropic trained successive LLMs, it became convinced that using books was the most cost-effective means to achieve a world-class LLM. During this time, however, Anthropic became 'not so gung ho about' training on pirated books 'for legal reasons' (Opp. Exh. 19). It kept them anyway."). And Plaintiffs agree to limit these requests and topics to the models that were trained on shadow-library data.

For its part, Anthropic argues that these requests and topics seek irrelevant information. Factually, Anthropic asserts that no publicly available Claude model was trained on text from the *Bartz* class repositories. It asserts that such text was only used for early research models that were not released publicly.

The Court **DENIES** this portion of Plaintiffs' motion to compel against Anthropic for the same reasons it denied the similar motion against OpenAI. Plaintiffs' motion assumes away differences in product development in contending that the literary generation capabilities of Anthropic's models that were trained on shadow libraries is relevant to showing the effect that using Plaintiffs' works for AI training has on the potential market. The evidence Plaintiffs seek could have nothing to do with Meta's fair use defense. Further, the requests and topics are

unreasonably broad in light of the questionable showing of relevance.

2. Licensing Evidence

Plaintiffs move to compel on the following document request: "3. All Documents and Communications Concerning Your licensing and efforts to license books and other content for AI training." They similarly move to compel on the following Rule 30(b)(6) topic: "3. Your efforts to license books and other content for AI training."

This dispute is the same as the one with OpenAI, and the Court's ruling is the same.

3. Torrenting Evidence

Plaintiffs move to compel on the following document requests:

> 4. Documents and Communications sufficient to identify each torrenting application, torrent software version, implementation, library, and/or Torrent Client You used to obtain Data from Online Databases from January 1, 2019 to the present.
>
> 5. All Documents and Communications, including configuration files, settings registries, and similar records, Concerning Your efforts to prevent, limit, or mitigate uploading, reuploading, seeding, peer exchange, leeching, or otherwise making Data available to other users during or after the Torrenting process, including by altering configurations and settings of Torrent Clients, devices, servers, virtual machines, networks, cloud environments, or storage systems.

They also move to compel on the following Rule 30(b)(6) topics:

> 4. The identity, versions, configuration, and settings of each Torrent Client You used to obtain Data from Online Databases from January 1, 2019 to the present.
>
> 5. Your efforts to prevent, limit, or mitigate uploading, re-uploading, seeding, peer exchange, or otherwise making Data available to other users during or after the Torrenting process.

This dispute is again the same as the one with OpenAI, and the Court's ruling is the same.

C. Conclusion

Plaintiffs' motions to compel against OpenAI and Anthropic are **DENIED**.

IT IS SO ORDERED.

Dated: September 10, 2026

THOMAS S. HIXSON
United States Magistrate Judge