UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

RAW STORY MEDIA, INC., ALTERNET MEDIA, INC.,

Plaintiffs,

v.

OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, and OPENAI HOLDINGS, LLC,

Defendants.

No. 1:24-cv-01514-CM

PLAINTIFFS' MEMORANDUM OF LAW IN OPPOSITION TO DEFENDANTS' MOTION TO DISMISS

TABLE OF CONTENTS

Cases

Statutes

Other Authorities

Rules

I. INTRODUCTION

Plaintiffs Raw Story and AlterNet have alleged that the OpenAI Defendants took a vast number of Plaintiffs' news articles without Plaintiffs' permission, removed their copyright management information, and used them to train ChatGPT products that incorporate and even completely regurgitate works on which they were trained. OpenAI's ChatGPT products, now worth billions of dollars, could not have been created without this training process—a process for which Defendants never requested, much less received, permission.

Defendants' motion contends that Plaintiffs are not entitled to their day in court on their Digital Millennium Copyright Act claims because they lack standing, have not adequately put Defendants on notice of their claims, and have not shown that Defendants acted with the required scienter. Not unlike the so-called "hallucinations" to which their products are sometimes prone, Defendants ignore or mischaracterize both the applicable law and the allegations of the Complaint. *See, e.g.*, Molly Bohannon, Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions (Forbes June 8, 2023), https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=34a922fa7c7f.

On standing, Defendants argue that Plaintiffs must allege specific works that ChatGPT disseminated to the public. But DMCA standing does not require dissemination. As with copyright infringement (the closest historical analogue to the DMCA), the injury is the interference with a plaintiff's right to exclude others from using its copyrighted works irrespective of dissemination. Plaintiffs have alleged just that.

Nor does the complaint fall short from a notice perspective. Having hidden much of the content of their training sets from public view, Defendants seek to leverage their own secrecy as a basis for dismissal, but their argument demands far more of Plaintiffs' Complaint than the pleading stage requires and ignores many of the Complaint's allegations. Based on allegations derived from the analysis of Plaintiffs' AI data scientist expert, Defendants are sufficiently on notice that Plaintiffs' claims are based on the works published on Plaintiffs' websites and placed into Defendants' training sets with author, title, and copyright information removed, and Defendants cannot and do not claim ignorance of the contents of their own training sets. Defendants also argue, disingenuously and incorrectly, that Plaintiffs were supposedly required to identify specific instances of regurgitation of their own works, but when the New York Times made exactly those allegations against them, Defendants accused them of nothing less than computer hacking. Memorandum of Law in Support of OpenAI Defendants' Motion to Dismiss, 2, *The New York Times Company v. Microsoft Corp.*, No. 23-cv-11195 (S.D.N.Y. Feb. 26, 2024).

Plaintiffs have adequately pled scienter as well. The Second Circuit allows for lenient scienter pleading, but even if a greater level of detail were required, Plaintiffs have pled abundant facts—well beyond the bare-bones pleadings that led to dismissals in some out-of-jurisdiction DMCA cases—including that ChatGPT plagiarizes substantial content. Defendants' contrary arguments rest on supposed pleading rules this District has expressly rejected.

In sum, Plaintiffs' claims are sufficiently pled to overcome a motion to dismiss. Defendants' motion should be denied.

II. BACKGROUND

A. Plaintiffs publish thousands of news articles online that contain CMI.

Plaintiffs Raw Story Media, Inc. and AlterNet Meda, Inc. are award-winning news organizations. Compl. ¶¶ 8-14. Their articles (created by professional authors who create media content with the expectation of earning a living) are published on the internet. *Id.* ¶ 31; *see also* rawstory.com; alternet.org. At the time of publication, their articles are conveyed with author, title, and copyright information. Compl. ¶ 31.

B. Defendants intentionally remove CMI from Plaintiffs' news articles.

Generative AI systems and large language models, including ChatGPT, are trained on works created by humans. Compl. ¶¶ 4-5. Once trained, an LLM is able to provide responses to user prompts. *Id.* ¶ 33. These responses sometimes mimic material from the works on which they are trained, and sometimes even "regurgitate" those works entirely for its own (paying) customers. *Id.* ¶¶ 34-36. When that happens, ChatGPT generally does not provide the author, title, and copyright information contained in the original version of the work. *Id.* ¶ 39.

Beginning with GPT-4, Defendants have hidden from the public and from copyright owners the precise contents of the training sets on which their products are built. *Id.* ¶ 28. But information exists about prior ChatGPT training sets, and that information shows that Defendants have trained their products on thousands of Plaintiffs' copyright-protected news articles. *Id.* ¶ 37. In particular, Defendant trained ChatGPT using training sets called WebText and WebText2—sets created by OpenAI that are collections of links posted on the website Reddit. *Id.* ¶¶ 29-30. Defendants also created training sets derived from a repository called Common Crawl, which is a vast "scrape of most of the internet" created by a third party. *Id.*

Defendants have not published the contents of WebText, WebText2, or their training sets derived from Common Crawl. *Id.* ¶ 28. But various public sources have recreated approximations of these datasets. *Id.* ¶ 37. And in those approximations, thousands of Plaintiffs' articles appear without the copyright management information with which Plaintiffs conveyed them to the public. *Id.* There is only one plausible explanation for this: Defendants intentionally removed the CMI. After all, given the nature of LLM training, if ChatGPT had been trained on works that included CMI, it would have learned to output CMI. *Id.* ¶ 38.

Likewise, Defendants knew, or had reasonable grounds to know, that their removal of Plaintiffs' CMI in their training sets would likely induce, enable, facilitate, or conceal infringement by both themselves and ChatGPT users. Indeed, after the tortious acts giving rise to this lawsuit were committed, Defendants have recently not only created tools to allow copyright owners to block their work from being incorporated into training sets, but they also reached licensing deals with some media organizations to pay them for the content they used for training, suggesting (at least in the light most favorable to Plaintiffs) that they know that copying journalists' works is likely infringement. *Id.* ¶¶ 57, 58. The removal of CMI, in turn, furthers and conceals Defendants' infringement at least by preventing ChatGPT users from knowing that outputs are based on copyright-protected works of journalism. *Id.* ¶ 46. It also furthers ChatGPT users' infringement at least by encouraging them to distribute outputs the users do not know are infringing. *Id.* ¶¶ 44-45. And it facilitates Defendants' large-scale copying and use of copyright-protected material in their training sets by avoiding the problems for their products that would arise if Defendants had included CMI. *Id.* ¶¶ 33, 38, 46.

III. LEGAL STANDARDS

Rule 8 requires only "a short and plain statement of the claim showing that the pleader is entitled to relief." Fed. R. Civ. P. 8(a)(2). On a motion to dismiss under Rule 12(b)(1) for lack of standing, the court's resolution depends on whether the motion is facial—based solely on the allegations in the complaint—or fact-based. *Sonterra Cap. Master Fund Ltd. v. UBS AG*, 954 F.3d 529, 533 (2d Cir. 2020). When the motion is facial, as Defendants' motions are here, the court must "accept[] as true all material factual allegations of the complaint, and draw[] all reasonable inferences in favor of the plaintiff." *Id.* (cleaned up). In these cases, "the plaintiff has no evidentiary burden." *Id.*

Likewise, under Rule 12(b)(6), the court must "accept[] all factual allegations as true and draw[] all reasonable inferences in favor of the plaintiff." *Sierra Club v. Con-Strux, LLC*, 911 F.3d 85, 88 (2d Cir. 2018). The court must deny the motion if the complaint "contain[s] sufficient factual matter, accepted as true, to 'state a claim that is plausible on its face.'" *Id.* (quoting *Ashcroft v. Iqbal*, 556 U.S. 662, 678, (2009)). And because the purpose of a complaint is to "give the defendant fair notice of what the ... claim is and the grounds upon which it rests," a plaintiff need not plead "specific facts" in order to overcome a motion to dismiss. *Erickson v. Pardus*, 551 U.S. 89, 93 (2007) (cleaned up) (quoting *Bell Atl. Corp. v. Twombly*, 550 U.S. 544, 545 (2007)). Defendants may dispute the accuracy of Plaintiffs' allegations or question Plaintiffs' ability to prove them, but those are issues for trial, not the pleading stage. *See Giuffre v. Dershowitz,* 410 F. Supp. 3d 564, 577 (S.D.N.Y. 2019) ("[P]laintiffs can put their allegations out to the world and must only plead them, not prove them, at the motion to dismiss stage.").

IV. ARGUMENT

A. Plaintiffs have Article III standing.

Article III standing requires, *inter alia*, a "concrete and particularized" injury. *Spokeo, Inc. v. Robins*, 578 U.S. 330, 339 (2016). Plaintiffs plausibly allege both.

1. Plaintiffs' injuries are concrete.

An injury is concrete if it has a "close historical or common-law analogue." *TransUnion LLC v. Ramirez*, 594 U.S. 413, 424 (2021). The analogy need not be an "exact duplicate." *Id.* at 433. Instead, "some relationship to a well-established common-law analog" will do. *Bohnak v. Marsh & McLennan Cos., Inc.*, 79 F.4th 276, 285 (2d Cir. 2023); *see also Saba Cap. Cef Opportunities 1, Ltd. v. Nuveen Floating Rate Income Fund*, 88 F.4th 103, 115-16 (2d Cir. 2023) (holding that dilution of voting shares is analogous to a common-law "property-based injury"). In deciding concreteness, "[c]ourts must afford due respect to Congress's decision to impose a statutory prohibition or obligation on a defendant, and to grant a plaintiff a cause of action to sue over the defendant's violation of that statutory prohibition or obligation." *TransUnion*, 594 U.S. at 425. Both "tangible" and "intangible" harms are concrete. *See id.* at 424-425.

The unlawful removal of CMI from a copyright-protected work—Plaintiffs' claim here—is analogous to copyright infringement. Congress evidently saw the two as analogous: it called the DMCA an Act "[t]o amend title 17, United States Code," which exclusively concerns copyright. Pub. L. 105-304, 112 Stat. 2860 (1998). It did so because CMI protects the integrity of copyrighted works. *See* S. Rep. 105-190 at 16 (1998). Further recognizing the analogy, Congress provided similar remedies for DMCA violations as it long had for copyright infringement: in both, the plaintiff can choose between actual damages and profits on the one hand, and statutory damages on the other. *Compare* 17 U.S.C. §§ 504(b), (c) (copyright infringement) *with* 17 U.S.C. § 1203(c) (DMCA violations)). While not dispositive, Congress's view on the matter is entitled to considerable weight. *See Spokeo*, 578 U.S. at 341 ("[B]ecause Congress is well positioned to identify intangible harms that meet minimum Article III requirements, its judgment is also instructive and important.").

The analogy between copyright infringement and DMCA violations also follows from first principles: both the Copyright Act and the DMCA protect similar rights involving copyright-protected works. The Copyright Act protects certain exclusive rights, such as the rights to reproduce the work and prepare derivative works. *See* 17 U.S.C. § 106 (listing exclusive rights). The DMCA grants copyright owners similar rights. In particular, the protection against removing or altering CMI, 17 U.S.C. § 1202(b)(1), is analogous to the rights to reproduce the works and prepare derivative ones, 17 U.S.C. §§ 106(1), (2): both grant the copyright owner the sole prerogative to decide how future iterations of the work may differ from the version the owner published.

Given this analogy, Plaintiffs have alleged a concrete injury: Defendants' interference with their exclusive right to control their copyrighted works by removing CMI from them. *See* Compl.

¶¶ 47-58. For copyright infringement, courts have never required more. *See Fox Film Corp. v. Doyal*, 286 U.S. 123, 127 (1932) (describing copyright owner's right as one simply to "exclude others from using his property"). This also accords with the common law, which recognizes interference with property, without more, as a concrete injury. *See* Restatement (Second) of Torts § 163 ("One who intentionally enters land in the possession of another is subject to liability to the possessor for a trespass, although his presence on the land causes no harm to the land, its possessor, or to any thing or person in whose security the possessor has a legally protected interest."). Given the infringement-DMCA analogy, the same holds for the latter: the unlawful removal of CMI from a copyrighted work is a concrete injury.

Defendants would impose two more conditions for standing. First, they would require that the defendant disseminate the plaintiff's works. *See* Mot. at 5-7. But copyright infringement—the relevant historical analogue—has never required dissemination. This has been true from the 1790 Copyright Act, passed by the first Congress, to the present version of the law. *See* Act of May 31, 1790, ch. 15, § 2 (imposing liability on anyone who "shall print, reprint, publish, or import" a copyrighted work); *A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1014 (9th Cir. 2001) (holding that downloading of files containing copyrighted music violates the reproduction right). Given the close analogy between copyright infringement and DMCA violations, nothing justifies treating DMCA violations differently.

Defendants contend that *TransUnion* supports a dissemination requirement. Mot. at 5-7. But it is far afield. The *TransUnion* plaintiffs alleged that a credit reporting agency did not keep accurate credit files. *See TransUnion*, 594 U.S. at 421. Likening their injury to one at common law, the plaintiffs chose defamation. *See id.* at 432. The Court held that the analogy justified a finding of injury for plaintiffs whose credit files were disseminated to third parties, but not for those whose files were not. *See id.* at 433. It reached that conclusion because defamation requires publication. *See id.* at 434. Thus, its dissemination requirement was an artifact of the plaintiffs' chosen analogy to a historical injury that requires it. That has no bearing on a case like this, where Plaintiffs analogize to a different historical injury—copyright infringement—that does not.

Second, Defendants would require economic harm. *See* Mot. at 7. But standing does not require this. *See TransUnion*, 594 U.S. at 425. And neither, historically, has copyright infringement—the relevant analogy. This is clear from the 1790 version of the Copyright Act which granted statutory damages of 50 cents per infringing page without any further showing. *See* Act of May 31, 1790, ch. 15, § 2. The same rule has persisted, with the Supreme Court making clear long ago that liability may lie "[e]ven for uninjurious and unprofitable invasions of copyright." *F. W. Woolworth Co. v. Contemp. Arts*, 344 U.S. 228, 233 (1952); *see also Jewell-La Salle Realty Co. v. Buck*, 283 U.S. 202, 208 (1931) (construing Copyright Act to mandate minimum statutory damages of $10 per performance even if "there is no showing as to actual loss"); 17 U.S.C. § 504(c)(1) (providing for statutory damages for copyright infringement without regard to economic loss). Given the analogy between DMCA violations and copyright infringement, the same result follows: DMCA violations do not require economic harm. Defendants cite no contrary authority.

2. Plaintiffs' injuries are particularized.

"For an injury to be 'particularized,' it 'must affect the plaintiff in a personal and individual way.'" *Spokeo*, 578 U.S. at 339 (quoting *Lujan v. Defs. of Wildlife*, 504 U.S. 555, 560 n.1 (1992)). Plaintiffs met that requirement by alleging that Defendants removed CMI from ***their*** copyright-protected news articles. *See* Compl. ¶¶ 47-58. If such removal constitutes an Article III injury—and it does for the reasons just given—then Plaintiffs have alleged that they suffered them in a personal and individual way.

Defendants resist this conclusion, relying on *Doe 1 v. Github*, 672 F. Supp. 3d 837 (N.D. Cal. 2023). That reliance is misplaced. *Doe 1* held that the plaintiffs did not identify a particularized injury sufficient to confer standing for damages because they did not allege dissemination of their own works (though, as discussed in Section IV.A.3, *infra*, it did find standing for an injunction). *See id.* at 850. But like *TransUnion*, *Doe 1* required dissemination only because of how the plaintiffs defined their injury: as a violation of their licenses, which prohibited dissemination without CMI. *See id.* Because the plaintiffs did not allege CMI-less disseminations of their own works, they "do not allege that they themselves have suffered the injury they describe," and thus failed the particularity requirement. *Id.* The Court offered no view on the analogy to copyright infringement because it was not asked to. Its holding on standing for damages is therefore inapposite. Plaintiffs have standing to pursue damages for Defendants' removal of Plaintiffs' CMI.

3. Even on Defendants' theory, Plaintiffs have standing to seek an injunction.

Plaintiffs seek both damages for Defendants' past CMI removal and an injunction requiring Defendants to remove their articles from their training sets. *See* Compl. at 11. Defendants do not directly dispute Plaintiffs' standing to seek injunctive relief, and both of Defendants' standing cases, *TransUnion* and *Doe 1*, only rejected plaintiffs' standing to seek damages. Even if correct, however, Defendants' arguments against standing for damages would not vitiate Plaintiffs' standing to seek an injunction, as the outcome of the standing analysis may differ between these forms of relief based on the same underlying facts. *See TransUnion*, 594 U.S. at 431 (holding that "plaintiffs must demonstrate standing for each claim that they press and for each form of relief that they seek (for example, injunctive relief and damages"); *id.* at 436-37 (holding that risk of future harm can ground a claim for an injunction but not a claim for damages).

In any case, Plaintiffs have standing to seek an injunction even on Defendants' inaccurate theory of standing as allegedly requiring dissemination. In fact, the Court need look no further than Defendants' lead case, *Doe 1*, which found standing for injunctive relief even though the Plaintiffs did ***not*** allege dissemination of their own works. *See Doe 1*, 672 F. Supp. 3d at 850-51. Specifically, it held that the plaintiffs had standing by alleging "a substantial risk that Defendants' programs will reproduce Plaintiffs' licensed code as output." *Id.* at 851. The plaintiffs had alleged that the programs were trained on their source code, that the programs sometimes reproduced well-known code (though not plaintiffs' own), and that one of the programs reproduces code "about 1% of the time." *Id.*

Plaintiffs here have easily cleared that bar. They allege that ChatGPT was trained on their copyrighted works, Compl. ¶ 37, that ChatGPT has reproduced copyrighted works of journalism, *id.* ¶¶ 34-35, and that "nearly 60% of the responses provided by Defendants' GPT-3.5 product in a study conducted by Copyleaks contained some form of plagiarized content, and over 45% contained text that was identical to pre-existing content," *id.* ¶ 5. This conveys a much greater risk than 1%. Thus, even if standing did require dissemination of Plaintiffs' own works, *but see* Section IV.A.1, *supra*, Plaintiffs have plausibly alleged facts to support standing for an injunction.

B. Plaintiffs were "injured" under section 1203(a).

Defendants argue that Plaintiffs "are not within the class of plaintiffs that Congress authorized to sue" for DMCA violations—a class Defendants do not even define—and thus that Plaintiffs did not suffer an injury under 17 U.S.C. § 1203(a) ("Any person injured by a violation of section 1201 or 1202 may bring a civil action."). *See* Mot. at 11 This argument fails for the same reason as the last: Plaintiffs suffered an Article III injury, and Defendants give no reason to believe that "injury" under section 1203(a) means anything different than it does under Article III.

Defendants further suggest that Plaintiffs must allege "harm flowing from ChatGPT's outputs," and thus, apparently, that section 1203(a)'s injury requirement requires dissemination of the plaintiff's works. Mot. at 12 But that argument belies the DMCA's text and structure and is therefore wrong. *See Food Mktg. Inst. v. Argus Leader Media*, 588 U.S. 427, 436 (2019) ("In statutory interpretation disputes, a court's proper starting point lies in a careful examination of the ordinary meaning and structure of the law itself."). Section 1202(b) creates three different violations: removing or altering CMI, distributing false or removed CMI, and distributing works knowing that CMI has been removed or altered. *See* 17 U.S.C. §§ 1202(b)(1), (2), (3). Congress chose to make dissemination an element of the second and third violations, but not the first—the one at issue here. Moreover, requiring dissemination for a section 1202(b)(1) removal claim would render that provision essentially duplicative of section 1202(b)(3): anyone who intentionally removed CMI from a work in violation section 1202(b)(1), and then disseminated the work, would necessarily have distributed it knowing that CMI had been removed, thus violating section 1202(b)(3). Congress thus determined that removal or alteration itself constitutes a harm under section 1202(b), regardless of whether or not the violator also disseminated the work.

Neither of the two cases cited by Defendants dictate a contrary conclusion. In *Steele v. Bongiovi*, 784 F. Supp. 2d 94, 97-98 (D. Mass. 2011), the plaintiff argued that the defendant's CMI removal caused him to lose a copyright suit, and the court held that the removal had no impact on the decision. And *Alan Ross Mach. Corp. v. Machinio Corp.*, No. 17-cv-3569, 2019 WL 1317664, at *4 (N.D. Ill. Mar. 22, 2019), held that the plaintiff asserted injury of "confusion in the marketplace" did not qualify because the defendant did not participate in that market. *Alan Ross Mach. Corp. v. Machinio Corp.*, No. 17-cv-3569, 2019 WL 1317664, at *4 (N.D. Ill. Mar. 22, 2019). Neither case suggests injury under the DMCA requires dissemination. It does not.

C. Plaintiffs state a claim under section 1202(b)(1).

A violation of section 1202(b)(1) requires "(1) the existence of CMI on the allegedly infringed work, (2) the removal or alteration of that information and (3) that the removal was intentional." *Fischer v. Forrest*, 968 F.3d 216, 223 (2d Cir. 2020). It also requires defendant to know, or have "reasonable grounds to know," that removal "will induce, enable, facilitate, or conceal" infringement. 17 U.S.C. § 1202(b). Plaintiffs sufficiently allege each element.

1. Plaintiffs adequately allege the existence and removal of CMI from their articles.

As to the first and second elements, Plaintiffs allege both the existence of CMI on their works and that Defendants removed it. *See* Compl. ¶¶ 31, 37-43. Defendants respond by arguing that Plaintiffs were required to, but did not, name the specific works from which Defendants removed their CMI. *See* Mot. at 12-14. The argument has no merit.

For starters, Plaintiffs have identified their works with the required level of detail in the context of this case. As Defendants admit, the purpose of identifying a work is to "give the defendants fair notice of the claims against them." Mot. at 13 (quoting *Dow Jones & Co., Inc. v. Int'l Sec. Exch., Inc.*, 451 F.3d 295, 307 (2d Cir. 2006)). Plaintiffs have alleged removal of CMI from their works contained in Defendants' training sets. *See* Compl. ¶ 41. And Defendants cannot seriously claim ignorance of what works those are. Indeed, OpenAI has "published a list of the top 1,000 domains present in WebText and their frequency," which notes that WebText—one of the training sets referenced in the Complaint—contains exactly 33,598 URLs (which it does not identify) from Raw Story's web domain and 23,183 from AlterNet's. *See* GPT-2 model card (Nov. 2019), https://github.com/openai/gpt-2/blob/master/model_card.md; Domains.txt (2019), https://github.com/openai/gpt-2/blob/master/domains.txt. This shows that it can isolate Plaintiffs' news articles in its training sets, which suffices to notify Defendants of the works at issue.[1]

Further, Plaintiffs cannot name all their works contained in Defendants' training sets precisely because Defendants have kept them secret. *See* Compl. ¶ 28. Requiring a plaintiff to name all its works in these circumstances—where, due to the defendants' efforts to conceal, only Defendants know (prior to discovery) what the works are—would perversely incentivize defendants to conceal their DMCA violations. And it would do so without advancing the principle animating the identification requirement: to provide defendants notice of the claims against them. If Defendants know the works, they know the claims.

In none of Defendants' laundry list of cases did the plaintiff's allegations, combined with information in the defendant's sole possession, enable the defendants to identify the works at issue. For instance, in this District's first case to require identification, the complaint alleged that the defendant infringed a work called "The Skyrider," but the plaintiff owned two works containing the word "Skyrider," and it was unclear to which work the complaint referred. *Calloway v. Marvel Ent. Grp., a Div. of Cadence Indus. Corp.*, No. 82-cv-8697, 1983 WL 1141, at *3-*4 (S.D.N.Y. June 30, 1983). Plaintiffs' Complaint contains no such ambiguity.

In Defendants' other cases that found the works insufficiently described—only one of which involved the DMCA—plaintiffs not only did not specifically name the works but (unlike here) also did not describe the works in a way that enabled defendants to identify them. *See Free*

[1] Though not specifically alleged in the Complaint, Plaintiffs' claims only relate to articles published on their own web domains, rawstory.com and alternet.org, which, for reasons just discussed, Defendants are able to identify. And to the extent that the copyrights are owned by their freelance authors, this issue can easily be addressed through discovery—in fact, it will likely be fully resolved through responses to requests for production Defendants have already propounded, in which they seek documents showing Plaintiffs' ownership of their works. If the Court prefers, Plaintiffs could replead to address this.

Speech Sys., LLC v. Menzel, 390 F. Supp. 3d 1162, 1175 (N.D. Cal. 2019) (plaintiff alleged CMI removal "without providing any facts to identify which photographs had CMI removed"); *see also Palmer Kane LLC v. Scholastic Corp.*, No. 12-cv-3890, 2013 WL 709276, at *3 (S.D.N.Y. Feb. 27, 2013) (plaintiff provided list of some works and indicated without elaboration that "this list is not exhaustive"); *Wolo Mfg. Corp. v. ABC Corp.*, 349 F. Supp. 3d 176, 201 (E.D.N.Y. 2018) (plaintiff described the allegedly infringed works simply as "Other Works"); *Cole v. John Wiley & Sons, Inc.*, No. 11-cv-2090, 2012 WL 3133520, at *12 (S.D.N.Y. Aug. 1, 2012) (for one defendant, plaintiff listed two works and alleged "that the claim is also intended to cover other, unidentified works"; for the others, plaintiff listed works but did not allege that defendants infringed any); *Joint Stock Co. Channel One Russia Worldwide v. Infomir LLC*, No. 16-cv-1318, 2017 WL 696126, at *14 (S.D.N.Y. Feb. 15, 2017) (plaintiff did not specify representative sample of the works, which in that case's context prevented defendants from evaluating the other elements); *Sherwood 48 Assocs. v. Sony Corp. of Am.*, 76 F. App'x 389, 391 (2d Cir. 2003) (plaintiff failed to precisely describe trade dress).

The remainder of Defendants' cases did not dismiss claims for failing to identify the works. *See Dow Jones*, 451 F.3d at 307 (dismissing trademark claim where plaintiff identified the marks but did not provide "any factual allegations concerning the nature of the threatened use"); *Felix the Cat Prods., Inc. v. Cal. Clock Co.*, No. 04-cv-5714, 2007 WL 1032267, at *4 (S.D.N.Y. Mar. 30, 2007) (noting that plaintiff had identified the work but dismissing complaint because it failed to specify other details concerning the infringement); *Sitnet LLC v. Meta Platforms, Inc.*, No. 23-cv-6389, 2023 WL 6938283, at *1 (S.D.N.Y. Oct. 20, 2023) (adopting additional provisions for case management plan). So they do not apply here.

In sum, in the context of this case, Plaintiffs have done all they need to put Defendants on notice of their claims, especially since the claims are the same for all of the works: Defendants took copies of works published by Plaintiffs on the internet without permission, stripped away copyright management information, and used them to train their products with the requisite intent and knowledge. This suffices to allege the existence and removal of CMI from Plaintiffs' works.

2. Plaintiffs adequately allege scienter.

A section 1202(b)(1) plaintiff must allege that the defendant removed the CMI both intentionally and while knowing, or with reasonable grounds to know, "that it will induce, enable, facilitate, or conceal an infringement of any right under this title." 17 U.S.C. § 1202(b)(1). Courts sometimes call this requirement "double-scienter." *Mango v. BuzzFeed, Inc.*, 970 F.3d 167, 171 (2d Cir. 2020).

The "scienter" label is significant because "[t]he Second Circuit has stated that courts should be lenient in allowing scienter issues to survive motions to dismiss." *Aaberg v. Francesca's Collections, Inc.*, No. 17-cv-115, 2018 WL 1583037, at *9 (S.D.N.Y. Mar. 27, 2018) (citing *In re DDAVP Direct Purchaser Antitrust Litig.*, 585 F.3d 677, 693 (2d Cir. 2009)). On that basis, courts in this District have allowed DMCA claims to proceed even on "sparse" allegations of scienter, including where plaintiff alleged only that defendant "'intentionally and knowingly removed copyright information identifying Plaintiff as the photographer of the Photograph'" and that the "'removal of said [CMI] was made by Defendants intentionally, knowingly, and with the intent to induce, enable, facilitate, or conceal their infringement of Plaintiff's copyright in the Photograph.'" *Hirsch v. CBS Broad. Inc.*, No. 17-cv-1860, 2017 WL 3393845, at *8 (S.D.N.Y. Aug. 4, 2017) (quoting complaint); *see also Devocean Jewelry LLC v. Associated Newspapers Ltd.*, No. 16-cv-2150, 2016 WL 6135662, at *2 (S.D.N.Y. Oct. 19, 2016) (denying motion to dismiss DMCA claims with "relatively sparse" allegations of scienter).

Plaintiffs here allege far more than the plaintiffs in these other cases, who themselves alleged enough to survive dismissal.

First, as to intent, Plaintiffs allege that "Defendants intentionally removed author, title, and copyright information from Plaintiff's copyrighted works in creating ChatGPT training sets." Compl. ¶ 42. This itself is enough. *See Hirsch*, 2017 WL 3393845, at *8. But Plaintiffs allege other facts too. For example, Plaintiffs allege that they published their works with specified CMI, while approximations of ChatGPT's training data contain copies of their works without CMI. *See* Compl. ¶¶ 31, 37. If the ChatGPT training data lacks CMI, it is at least plausible that the data's creators—Defendants—intentionally removed it to optimize the training process. Indeed, that is the ***only*** plausible explanation. How else would the CMI have gotten stripped away?

In disputing that Plaintiffs plausibly allege intent, Defendants simply ignore Plaintiffs' allegations about the public approximations of ChatGPT's training sets containing Plaintiffs' CMI-less works. They instead argue that Plaintiffs' theory rests on supposedly "speculat[ive]" inferences about the relation between ChatGPT's inputs and its outputs. Mot. at 15. Yet, as just shown, Plaintiffs plausibly allege intentional CMI removal without needing to resort to that relation. Even if it were required, though, Defendants' arguments fail on their own terms.

For one, Defendants assume that Plaintiffs can allege intent only if ChatGPT has regurgitated Plaintiffs' own works. *See* Mot at 15. That assumption is unsupported and incorrect. Plaintiffs allege that (1) ChatGPT's training sets include both Plaintiffs' and others' news articles, and (2) ChatGPT's outputs regurgitate copyright-protected works of journalism without CMI. *See* Compl. ¶¶ 4, 34. Absent any reason to believe that Defendants treat different news organizations' differently—which the Complaint does not allege, and which in fact there is not—then there is only one natural inference: CMI-less regurgitations of some news articles means removal of CMI from all news articles, including those owned by Plaintiffs. Plus, a DMCA claim does not require any actual infringement of a plaintiff's copyright*, see Murphy v. Millennium Radio Grp. LLC*, No. 08-cv-1743, 2015 WL 419884, at *5 (D.N.J. Jan. 30, 2015), so it does not require regurgitation of a plaintiff's works for that reason either. Thus, since Defendants know that regurgitation occurs, Plaintiffs have sufficiently alleged that the requisite intent and knowledge existed at the time the CMI was removed, regardless of the extent to which Defendants' products regurgitate Plaintiffs' works.

Defendants also argue that it is too speculative to infer that the absence of CMI from ChatGPT outputs results from an intentional choice to remove CMI from the training data. *See* Mot. at 15. But it is at least plausible that ChatGPT doesn't output CMI precisely because it wasn't trained on CMI. While other hypothetical explanations (of which, by the way, Defendants offer none) might conceivably be available, they can be posited, tested, and challenged during discovery. But regardless, Plaintiffs need not disprove them at the pleading stage. *See George & Co. LLC v. Target Corp.*, No. 21-cv-4254, 2022 WL 1407236, at *12 (E.D.N.Y. Jan. 27, 2022) (holding that a complaint need not "disprove an obvious alternative explanation" to satisfy *Twombly*) (cleaned up).

Defendants cite *Tremblay* v. *OpenAI, Inc.*, No. 23-cv-03223, 2024 WL 557720 (N.D. Cal. Feb. 12, 2024), which found intent lacking, but the allegations there are distinguishable. While *Tremblay* held insufficient the bare allegation that defendants removed CMI "by design," *id*. at *4, Plaintiffs in this case have alleged more, including the absence of CMI in the approximated datasets. Moreover, unlike here, the *Tremblay* plaintiffs included excerpts of ChatGPT outputs "that include multiple references to Plaintiffs' names," thus undermining the plaintiffs' own allegations that OpenAI had removed the CMI. *Id*. at *4. Plaintiffs' complaint contains no similar self-defeating allegations.

Defendants attempt to manufacture one by pointing to the allegation that outputs sometimes provide author and title. *See* Mot. at 17 (citing Compl. ¶ 39). But they misleadingly fail to cite the rest of the paragraph, namely, that outputs likely do so only "because other material used in a training set references the author or title in the text of such material (e.g., a Wikipedia article discussing the underlying works of journalism)." Compl. ¶ 39. Thus, when an output references author or title, it is ***not*** because Defendants left the CMI in Plaintiffs' articles. It is because the content of someone else's article happened to attribute the material to Plaintiffs.

In addition to adequately alleging intent, Plaintiffs also plausibly allege that Defendants removed their CMI knowing, or with reason to know, "that it will induce, enable, facilitate, or conceal an infringement." 17 U.S.C. § 1202(b). The relevant infringement can be a third party's or the defendant's. *See Mango*, 970 F.3d at 172. If the infringement is a future one, it need not be certain; it only must be "likely." *Id*.

Plaintiffs have easily cleared this bar too. They allege not only that Defendants had the requisite scienter—which by itself is enough under *Hirsch*—but give further reasons why. For one, Defendants had reason to know that their CMI removal would conceal their own infringements: because any ChatGPT responses that incorporated or regurgitated Plaintiffs' works would lack CMI, those responses would not communicate to ChatGPT users that they infringed Plaintiffs' copyright. *See* Compl. ¶ 43. Defendants also had reason to know that removal would induce, enable, or facilitate infringement by ChatGPT users. That is because they promote ChatGPT as a tool that can be used to generate content for a future audience, and at least some users would be less likely to distribute ChatGPT's responses based on Plaintiffs' works if they knew those works were copyright-protected. *See id*. ¶¶ 44-45.

Moreover, the risks of future infringements are plainly likely, as the award-winning website, Copyleaks, found that "nearly 60% of the responses provided by Defendants' GPT-3.5 product in a study conducted by Copyleaks contained some form of plagiarized content, and over 45% contained text that was identical to pre-existing content." *Id.* ¶ 5. Further OpenAI has reached licensing deals with some media organizations and created tools "to allow copyright owners to block their work from being incorporated into training sets." *Id*. ¶¶ 57, 58. Viewed in the light most favorable to Plaintiffs, that evidences knowledge that the use in training, and output, of copyrighted-protected works constitutes infringement.

Defendants respond primarily by purporting to refute an argument Plaintiffs do not make: they say that because the training sets are secret, removing CMI from them does not conceal their own infringement from the public. *See* Mot. at 15-17. But that does not address Plaintiffs' actual scienter theory: Defendants knew or had reason to know that CMI removal during training results in CMI-less ***outputs*** that misinform the public about the source of the content Defendants' products provide to users.

Defendants also argue that Plaintiffs have not specifically identified such outputs—ironic, given their accusation that the New York Times "paid someone to hack OpenAI's products" by doing just that. Memorandum of Law in Support of OpenAI Defendants' Motion to Dismiss, 2, *The New York Times Company v. Microsoft Corp.*, No. 23-cv-11195 (S.D.N.Y. Feb. 26, 2024). But regardless, such allegations are not necessary at this stage, particularly given the well-pleaded allegation regarding the extent to which ChatGPT's outputs plagiarize online content. *See* Compl. ¶ 4. That plagiarism makes it likely that ChatGPT's outputs infringe Plaintiffs' copyrights and that Defendants at least had reason to know that removing Plaintiffs' CMI would conceal their infringement. Moreover, Plaintiffs have plausibly alleged that Defendants had reason to know that their removal of CMI enables and facilitates their large-scale copying and use of copyright-protected material in their training sets by avoiding the practical problems for their products that would arise if Defendants had included CMI. *Id.* ¶¶ 34, 39, 50.

Indeed, the only case Defendants cite for this supposed pleading requirement, *Stevens v. CoreLogic*, 899 F.3d 666, 674 (9th Cir. 2018), was decided on summary judgment. For that reason, later cases in the Ninth Circuit found it "inapposite" at the pleading stage, as resolution of scienter issues "is more suited to summary judgment." *Izmo, Inc. v. Roadster, Inc.*, No. 18-cv-06092, 2019 WL 13210561, at *4 (N.D. Cal. Mar. 26, 2019); *see also Doe 1*, 672 F. Supp. 3d at 858 (holding that *Stevens*, as a summary judgment case, does not alter the rule that, at the pleading stage, "mental conditions generally need not be alleged with specificity"). Far more apposite are this District's decisions in *Aaberg*, *Hirsch*, and *Devocean Jewelry*, all of which properly applied the Second Circuit's lenient scienter pleading standards and allowed the case to move forward. *See Aaberg*, 2018 WL 1583037, at *9; *Hirsch*, 2017 WL 3393845, at *8; *Devocean Jewelry*, 2016 WL 6135662, at *2. This Court should do the same.

V. CONCLUSION

Defendants' motions to dismiss are based on unrecognized and unsupported legal theories, inflated pleading standards, and disregard for many of the Complaint's well-pleaded allegations. The Court should therefore deny the motions. In the alternative, the Court should grant Plaintiffs leave to replead to correct any deficiencies the Court identifies. *See Cortec Indus., Inc. v. Sum Holding L.P.*, 949 F.2d 42, 48 (2d Cir. 1991) ("It is the usual practice upon granting a motion to dismiss to allow leave to replead.").

/s/ Matthew Topic

Jonathan Loevy (*pro hac vice*)
Michael Kanovitz (*pro hac vice*)
Lauren Carbajal (*pro hac vice*)
Stephen Stich Match (No. 5567854)
Matthew Topic (*pro hac vice*)

LOEVY & LOEVY
311 North Aberdeen, 3rd Floor
Chicago, IL 60607
312-243-5900 (p)
312-243-5902 (f)
jon@loevy.com
mike@loevy.com
carbajal@loevy.com
match@loevy.com
matt@loevy.com

Counsel for Plaintiff

Dated: May 13, 2024