United States District Court
Northern District of California

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SARAH ANDERSEN, et al.,

Plaintiffs,

v.

STABILITY AI LTD., et al.,

Defendants.

Case No. 23-cv-00201-WHO (LJC)

ORDER DENYING PLAINTIFFS' REQUEST FOR NON-LAION TRAINING DATASETS

Re: Dkt. No. 307

The parties dispute whether Defendant Midjourney must produce all datasets it used to train its generative AI models or produce only datasets sourced from LAION.[1] Plaintiffs argue that the production of all datasets is warranted and these "datasets are among the most crucial evidence in this case." ECF No. 307 at 2. Midjourney objects that non-LAION datasets are irrelevant to Plaintiffs' claims and their production would be overly burdensome. The Court sides with Midjourney and denies Plaintiffs' request for an order compelling production of the non-LAION datasets.

Rule 26(b)(1) establishes that parties "may obtain discovery regarding any nonprivileged matter that is *relevant to any party's claim or defense* and proportional to the needs of the case[.]" (emphasis added). Plaintiffs contend that the non-LAION datasets are relevant to their claims because, at heart, their claims are about whether Midjourney copied registered artwork, "whether the provenance of that copy was LAION or otherwise." ECF No. 307 at 2. This expansive view of their case is at odds with the scope of the operative complaint. Plaintiffs allege that their

[1] LAION ("Large-Scale Artificial Intelligence Open Network") is an organization that makes large-scale machine learning models and datasets available to the public. SAC ¶ 57-58. As relevant here, LAION released: LAION-400M, a dataset of URLs and metadata of 400 million training images; LAION-5B, a dataset of URLs and metadata of 5.85 billion training images; and LAION-2B, a subset of images in LAION-5B that include English-language captions (collectively, the "LAION datasets"). *Id.* ¶¶ 59-67, 72.

artwork was included in the LAION datasets, and that Midjourney used the LAION datasets, as well as datasets from "private data partners," to train their image-generating AI models. ECF No. 238 (Second Am. Compl.) ¶¶ 71, 258. Their Second Amended Complaint asserts two sets of claims against Midjourney for 1) direct copyright infringement of the works contained in the LAION datasets (Counts Four and Five) and 2) Lanham Act claims for unauthorized commercial use of Plaintiffs' names (Count Seven) and vicarious trade dress infringement (Count Eight).[2]

Plaintiffs' Count Four, brought on behalf of the "LAION-400M Registered Plaintiffs and Damages Subclass,"[3] asserts that Midjourney directly infringed on copyrighted works contained in the LAION-400M dataset. *Id.* ¶ 271-75. Their Count Five, brought on behalf of the "LAION-5B Registered Plaintiffs and Damages Subclass,"[4] asserts that Midjourney directly infringed on copyrighted works contained in the LAION-5B dataset. *Id.* ¶ 276-82.[5] These copyright claims are explicitly limited to Midjourney's alleged copying from the LAION datasets. *See* ECF No. 223 at 19-21 (denying Midjourney's motion to dismiss Plaintiffs' copyright claims where Plaintiffs had plausibly pled that "their works were included in the LAION datasets"). Despite this, Plaintiffs first argue that obtaining non-LAION datasets is warranted because evidence that Midjourney copied Plaintiffs' registered works from other, non-LAION sources would "definitively" establish "a violation of copyright law[.]" ECF No. 307 at 2. This argument both puts the cart before the horse—whether using registered works to train an AI model is "a violation

---

[2] Plaintiffs' Second Amended Complaint also asserts claims under the Digital Millennium Copyright Act, but these claims were previously dismissed with prejudice. *See* ECF No. 223 at 13 n.13.

[3] LAION-400M Registered Plaintiffs "denotes the subset of Plaintiffs who hold registered copyrights" in artwork in the LAION-400M dataset. Second Am. Compl. ¶ 261. The LAION-400M Damages Subclass is defined as persons in the United States "that own a registered copyright in any work in the LAION-400M dataset that was used to train any version of an AI image product offered" by Defendants. *Id.* at 11.

[4] LAION-5B Registered Plaintiffs "denotes the subset of Plaintiffs who hold copyrights in" works in the LAION 5B dataset that were registered before this action commenced. Second Am. Compl. ¶ 213. The LAION-5B Damages Subclass is defined as all persons in the United States "that own a registered copyright in any work in the LAION-5B dataset that was used to train any version of an AI image product offered" by Defendants. *Id.* at 11.

[5] Plaintiffs' argument that they brought Counts Four and Five on behalf of the damages subclass who had registered artwork in the LAION datasets instead of the damages class who had copyright interests in any works that were used to train Defendants' AI models irrespective of training datasets was an accident as a result of "drafting inconsistenc[ies]" is unpersuasive. ECF No. 307 at 4 n.2.

of copyright law" is an open question being actively litigated in this and other generative AI cases—and does nothing to show that non-LAION datasets are relevant to Plaintiffs' copyright claims against Midjourney, which, as is amply clear in the Second Amended Complaint, are specifically limited to Midjourney's alleged copying of images contained in the LAION datasets. As Midjourney argues, Plaintiffs cannot use discovery "as a fishing expedition for evidence of claims that have not been properly pled" or potential future lawsuits. *Impinj, Inc. v. NXP USA, Inc.*, No. 19-cv-03161-YGR-AGT, 2022 WL 16586886, at *2 (N.D. Cal. Nov. 1, 2022) (citations omitted).

Plaintiffs' second argument why non-LAION datasets are relevant to this lawsuit is that they anticipate Midjourney will argue that "Plaintiffs' works comprise a very small fraction of its models['] datasets, and thus any infringement would be fair use." ECF No. 307 at 2. The fair use doctrine establishes that "the fair use of a copyrighted work…for purposes such as criticism, comment, news reporting, teaching…, scholarship, or research, is not an infringement of copyright." *Kadrey v. Meta Platforms, Inc.*, No. 23-cv-03417-VC, 2025 WL 1752484, at *3 (N.D. Cal. June 25, 2025) (quoting 17 U.S.C. § 107). While fair use is a "flexible concept," the Copyright Act lists four factors to determine whether use of a copyrighted work is fair: "(1) the purpose and character of the use…; (2) the nature of the copyrighted work; (3) the amount and sustainability of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work." *Id.* (quoting *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 598 (2023)); 17 U.S.C. § 107.

Plaintiffs argue that they need to have access to Midjourney's entire training corpus to rebut the third factor and show that "their works (or LAION in general)" is not such a "small proportion" of Midjourney's training data so that "any infringement is excused." ECF No. 307 at 2. But this misconstrues the fair use doctrine. Under the third factor, courts consider the amount of the copyrighted work the alleged copier uses relative to the "copyrighted work as a whole[.]" 17 U.S.C. § 107. The more of a copyrighted work an alleged copier uses and makes public, the less likely such use is fair, and vice versa. For example, a musical parody that samples a two-

second generic snippet of a copyrighted song has a stronger fair use argument than one that samples long, recognizable "portions of the original song[.]" *Kadrey*, 2025 WL 1752484, at *14. As applied to this case, the third factor would likely focus on the proportion of a registered artwork copied by Midjourney and produced as outputs by Midjourney's AI models compared to the registered artwork as a whole. *See id.*; *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 188 (2d Cir. 2024) ("The relevant consideration…[is] the amount of copyrighted material *made available to the public*.") (quotations omitted). Whether Plaintiffs' artworks comprise an insignificant portion of Midjourney's overall training data does not factor into this analysis. Moreover, it is not clear to the Court why Plaintiffs would need the entire corpus of training data rather than targeted information regarding the size and, potentially, sources of non-LAION data, which, as Defendants note, Plaintiffs could obtain "by way of a targeted interrogatory." ECF No. 307 at 6.

The Court accordingly finds that non-LAION datasets are not relevant to Plaintiffs' copyright claims against Midjourney. However, the Court notes that Plaintiffs' argument regarding Midjourney's anticipated fair use defense is, at this point, abstract. Plaintiffs may renew their request for non-LAION training datasets based on Midjourney's fair use defense only if, at a later stage in the case, they can concretely articulate how the content (rather than the overall size or sources) of non-LAION datasets would be relevant to rebutting Midjourney's fair use defense.

Plaintiffs assert two Lanham Act claims against Midjourney. Midjourney allegedly published a list of artists (the Name List), including many named Plaintiffs (the Name List Plaintiffs), whose styles its generative AI models could emulate. Second Am. Compl. ¶¶ 254-56. Plaintiffs' Count Seven, for false endorsement, asserts that Midjourney violated the Lanham Act by releasing the Name List, which "created a likelihood of confusion over whether the" Name List Plaintiffs endorsed or were affiliated with Midjourney's products. Second Am. Compl. ¶ 299. Plaintiffs' Count Eight, for vicarious trade dress infringement, alleges that Midjourney violated the Lanham Act by profiting from imitations of the Name List Plaintiffs' protectable artistic styles. *Id.* ¶¶ 312-17. Both of these claims hinge on whether Midjourney's use of the Name List Plaintiffs' names and styles was likely to cause confusion or mistake as to the affiliation between

the Name List Plaintiffs and Midjourney. *See* 15 U.S.C. 1125(a)(1)(A); *Art Attacks Ink, LLC v. MGA Ent. Inc.*, 581 F.3d 1138, 1145 (9th Cir. 2009).

Plaintiffs do not offer any argument as to how the non-LAION datasets could be necessary to their false endorsement claim, which turns on whether the Names List gave the false impression that the Names List Plaintiffs approved of or were associated with Midjourney. The content of Midjourney's' training data has no bearing on that claim. Plaintiffs argue that the non-LAION datasets are relevant to their vicarious trade dress claim, because "comparing whether there is a likelihood of confusion between Plaintiffs' works and Midjourney's outputs require comprehensive comparison of Plaintiffs' works with the entirety of what is in Midjourney's [entire] training corpus[.]" ECF No. 307 at 2-3. Comparing Plaintiffs' works and Midjourney's outputs will certainly be necessary, but, Plaintiffs' conclusory statements to the contrary, the undersigned does not see how Midjourney's training data bears on this comparison.

The undersigned accordingly finds that Plaintiffs have not established that non-LAION training data is relevant to their claims against Midjourney. *In re Glumetza Antitrust Litig.*, No. 19-cv-05822-WHA (RMI), 2020 WL 3498067, at *7 (N.D. Cal. June 29, 2020) (the party seeking discovery has the "initial burden of establishing that the request satisfies the relevancy requirements of Fed. R. Civ. P. Rule 26(b)(1)"). Because Plaintiffs have not demonstrated how this information would be relevant, ordering Midjourney to produce it would be disproportionate "to the needs of the case" and contrary to the limits imposed by Rule 26(b)(1). Plaintiffs' request for non-LAION training data is accordingly denied.

IT IS SO ORDERED.

Dated: June 27, 2025

LISA J. CISNEROS
United States Magistrate Judge