QUINN EMANUEL URQUHART & SULLIVAN, LLP
Sean S. Pak (SBN 219032)
seanpak@quinnemanuel.com
50 California Street, 22nd Floor
San Francisco, CA 94111
Telephone: (415) 875-6600
Facsimile: (415) 875-6700

Andrew H. Schapiro (admitted *pro hac vice*)
andrewschapiro@quinnemanuel.com
191 N. Wacker Drive, Suite 2700
Chicago, Illinois 60606
Telephone: (312) 705-7400
Facsimile: (312) 705-4001

Alex Spiro (admitted *pro hac vice*)
alexspiro@quinnemanuel.com
295 Fifth Avenue
New York, NY 10016
Telephone: (212) 849-7000
Facsimile: (212) 849-7100

Rachael L. McCracken (SBN 252660)
rachaelmccracken@quinnemanuel.com
865 South Figueroa Street, 10th Floor
Los Angeles, CA 90017
Telephone: (213) 443-3000
Facsimile: (213) 443-3100

Attorneys for Defendant, NVIDIA Corporation

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

OAKLAND DIVISION

ABDI NAZEMIAN, BRIAN KEENE, STEWART O'NAN, ANDRE DUBUS III and SUSAN ORLEAN, individually and on behalf of all others similarly situated,

Plaintiffs,

v.

NVIDIA CORPORATION,

Defendant.

Master File Case No. 4:24-cv-01454-JST (SK)

DEFENDANT'S RESPONSE TO MOTION FOR RELIEF FROM NONDISPOSITIVE PRETRIAL ORDER OF MAGISTRATE JUDGE

PRELIMINARY STATEMENT

Plaintiffs ask this Court to overturn the Magistrate Judge's decision which correctly limited discovery to the dataset on which Plaintiffs based their infringement claim—consistent with the decisions of every court to examine the issue. This case is about whether NVIDIA's use of the Books3 dataset to train its "NeMo Megatron" family of large language models ("LLMs") is copyright infringement or fair use. Plaintiffs focused their case on a single dataset (Books3)—which is a subset of a larger dataset called "The Pile." ECF 1 ("Compl.") ¶¶ 25-26, 29, 31 (identifying only Books3 as containing Plaintiffs' works). Plaintiffs did ***not*** allege that their works are found in any dataset other than Books3. Similarly, Plaintiffs alleged only that NVIDIA infringed their copyrights by using Books3 to train a single model family, the NeMo Megatron family of LLMs. *Id.* ¶¶ 23, 31, 34 ("To train the NeMo Megatron language models, NVIDIA copied The Pile dataset.").

Despite these narrow allegations, Plaintiffs have sought sweeping discovery into LLMs and datasets (which Plaintiffs call "shadow libraries") not identified in the Complaint and that Plaintiffs did not allege were trained on or contain copies of their works. ECF 162. Plaintiffs argued to the Magistrate Judge that discovery into models and datasets "beyond those named in the Complaint" might allow them to make other claims than what they pled in the Complaint. *Id.* at 1-4. The Magistrate Judge considered and correctly rejected these arguments, properly limiting discovery to (1) "the dataset that Plaintiffs know that Defendant used and that contains Plaintiffs' copyrighted books"—the Pile—and (2) LLMs "in the Nemo Megatron family that exist and that trained on the Pile dataset." ECF 168 (the "Order") at 2-3. Plaintiffs challenge only the decision on datasets.

Plaintiffs seek to overturn the Magistrate Judge's decision and compel NVIDIA to run Search Term 1 to fish for information untethered to the allegations in their Complaint. They acknowledge as much in a footnote. *See* Mot. at 1 n.2 ("Should discovery show Nvidia used other shadow libraries to train its LLMs, Plaintiffs may seek to expand the scope of models through a discovery motion or leave to amend."). Every court to examine requests like Plaintiffs' for discovery into datasets untethered to the operative complaint or the plaintiffs' works has rightly denied those requests, as the Order did. Because the Magistrate Judge committed no error, let alone clear error, the Court should deny the Motion.

ARGUMENT

A magistrate judge's non-dispositive order may only be overturned if there is "a definite and firm conviction that a mistake has been committed." *Perry v. Schwarzenegger*, 268 F.R.D. 344, 348 (N.D. Cal. 2010). "The court should not disturb the magistrate's relevance determination except where it is based on an erroneous conclusion of law or where the record contains no evidence on which the magistrate rationally could have based that decision." *Id.* (citation omitted). Plaintiffs here fail to show the Order is "clearly erroneous or contrary to law." Fed. R. Civ. P. 72(a).

Plaintiffs' Direct Infringement Claim Is Limited To The Pile. The Complaint alleges "NVIDIA copied The Pile dataset," which "includes the Books3 dataset, which includes the Infringed Works," and that NVIDIA "made multiple copies of the Books3 dataset while training the NeMo Megatron models." Compl. ¶ 34. Nowhere do Plaintiffs allege their asserted works can be found in any dataset other than The Pile or that NVIDIA used any dataset other than The Pile to train the NeMo Megatron LLMs.[1] The Complaint's sole generic reference to "shadow libraries" does not allege NVIDIA used them or that they contain Plaintiffs' works and does not purport to base any cause of action on them. *Id.* ¶ 27 (noting "shadow libraries" "have long been of interest to the AI-training community because they host and distribute vast quantities of unlicensed copyrighted material"). This passing reference, without more, cannot make these datasets relevant to Plaintiffs' infringement claim. Further, Plaintiffs' suggestion that NVIDIA did not deny having "knowledge (or possession)" of these "shadow libraries" in its answer is misleading, at best. Mot. at 3. The Complaint did not allege any such knowledge or possession, so it was not addressed in the answer. Compl. ¶ 27.

There is no precedent for the untethered discovery Plaintiffs seek into unpled datasets. Like the Magistrate Judge's Order, every case that has considered discovery requests for datasets not identified in the complaints' infringement allegations has rejected them. In *Anderson v. Stability*

[1] Shortly before the hearing before the Magistrate Judge, Plaintiffs asserted for the first time that their works can be found in LibGen, Anna's Archive, and Z-Library. Plaintiffs raise the same argument in a single sentence and footnote. Mot. at 2 n.3 (citing a URL to a third-party website). Plaintiffs do not dispute that the Complaint lacks allegations regarding these datasets, so their new assertion is irrelevant and should be disregarded. In any event, Plaintiffs' lone citation to a third-party website, without any allegations directed at ***NVIDIA***, cannot justify this new, unpled theory.

AI Ltd., the plaintiffs' complaint asserted infringement of works "contained in the LAION datasets," but plaintiffs nonetheless sought discovery into all datasets used to train the accused AI models. No. 23-cv-00201-WHO, ECF 307 at 3-5, ECF 314 at 2. The court denied the plaintiffs' broad request, finding that (1) their "expansive view of their case [wa]s at odds with the scope of the operative complaint," and (2) discovery into non-LAION datasets—which were not alleged to include Plaintiffs' copyrighted works—was a "fishing expedition" "disproportionate 'to the needs of the case' and contrary to the limits imposed by Rule 26(b)(1)." *Id.* at 3-5 (citation omitted).

Similarly, in *In re Mosaic LLM*, the court granted a motion to dismiss and denied discovery on similar facts, even where Plaintiffs later attempted to add allegations directed to additional datasets, finding "Plaintiffs do not allege facts that could establish that the [AI] models are actually trained on any shadow library websites, let alone those that contain Plaintiffs' works." 2025 WL 2402677, at *2-3 (N.D. Cal. Aug. 19, 2025); *In re Mosaic LLM Litig.*, No. 24-CV-01451-CRB (N.D. Cal. Aug. 20, 2025), ECF 163 (denying discovery requests for "all [AI model] training data" because "[t]o the extent that the disputed discovery requests relate to Plaintiffs' surviving claims, Plaintiffs have not demonstrated that their requests are proportional to the needs of this case going forward"); *In re Google Generative AI Copyright Litig.*, 2025 WL 2624885, at *7 (N.D. Cal. Sept. 11, 2025) (granting motion to dismiss because "Plaintiffs do not allege that any of their works were included in training datasets used to develop these models"). Likewise, in *Kadrey v. Meta Platforms, Inc.,* the court denied a motion to compel Meta to "identify all copies it made of copyrighted works, including but not limited to Plaintiffs' works." No. 3:23-cv-03417 (N.D. Cal. Dec. 20, 2024), ECF 401 at 2. The court explained that the case "is about the use of copyrighted materials to train the Llama models, not all copyright infringement committed by Meta" and ***"datasets not used to train the Llama models are not relevant or proportional to the case."*** *Id.* (emphasis added). Here, too, Plaintiffs do not even allege NVIDIA used any datasets other than The Pile to train the NeMo Megatron LLMs.

Plaintiffs' reliance on *Bartz v. Anthropic* is misplaced. In *Bartz*, unlike here, the plaintiffs alleged their copyrighted works were included in ***multiple*** datasets. No. 3:24-cv-05417 (N.D. Cal.), ECF 1 ¶¶ 56-58 ("Pirated copies of [plaintiffs'] work are available online through websites like

LibGen and Bibliotek."). The plaintiffs further alleged that Anthropic copied their works, "including ***but not limited to*** the books contained in Books3." *Id.* ¶ 43 (emphasis added). Similarly, in *Tremblay v. OpenAI*, the plaintiffs alleged that OpenAI's ChatGPT LLMs were trained on a dataset "estimated to contain about 294,000 [book] titles," and that the "only 'internet-based books corpora' that have ever offered that much material are notorious 'shadow library' websites like Library Gensis (aka LibGen), Z-Library (AKA B-ok), Sci-Hub, and Bibliotik." No. 3:23-cv-03223 (N.D. Cal.), ECF 1 ¶ 34. Plaintiffs made no such allegations here. *Compare id. with* Compl. ¶ 27. The fact that defendants ***in unrelated cases*** used datasets that Plaintiffs here neither allege contain their works, nor allege were used to train the NeMo Megatron LLMs, is irrelevant to Plaintiffs' claims against NVIDIA. It cannot render the Order clearly erroneous.[2]

Plaintiffs' Proposed Class Is Limited By Their Own Allegations, Not The Order. The Complaint alleges NVIDIA used The Pile—and only The Pile—to train four NeMo Megatron LLMs. And Plaintiffs seek to represent a class of all persons who own a "copyright in any work that was used as training data for the NeMo Megatron [LLMs]."[3] Compl. ¶ 39. But Plaintiffs do not allege NVIDIA used any dataset other than The Pile to train the NeMo Megatron LLMs at issue in this case. While Plaintiffs claim they need discovery due to information asymmetry—an argument the Magistrate Judge rejected—this cannot justify a fishing expedition untethered to any allegation in the Complaint. Order at 3 ("This is an inherent problem in any litigation: a plaintiff might not be able to discover wrongdoing by a defendant. That lack of knowledge does not justify a discovery request without bounds."); *Impinj, Inc. v. NXP USA, Inc.*, 2022 WL 16586886, at *2 (N.D. Cal.

---

[2] Plaintiffs' attempt to distinguish *Stability AI* and *Meta* lacks merit. Mot. at 4 n.5. That the plaintiffs sought the production of the datasets themselves and not communications about those datasets is immaterial. Neither decision turned on the burden of producing datasets compared to reviewing and producing custodial documents. Moreover, Plaintiffs here seek more than just "discussions of other pirated libraries." *See, e.g.*, Ex. 1 (RFP 13 seeking "all datasets, datasources, and documentation detailing the content and composition of the Training Data, including but not limited to The Pile and Books3 Dataset."); Ex. 2 (Interrogatory 12 asking to "[l]ist all datasets that have been permitted to be used as Training Data for any NVIDIA Language Model").

[3] The *Bartz* plaintiffs sought to represent a much broader class. *See* No. 3:24-cv-05417 (N.D. Cal.), ECF 1 ¶ 59 (defining the class as all persons owning copyrighted works that "were or are used by Defendant in LLM training, research, or development, including but not limited to training Defendant's Claude family of models").

Nov. 1, 2022) ("discovery cannot be used as a fishing expedition for evidence of claims that have not been properly pled") (citation omitted); *In re Mosaic*, 2025 WL 2402677, at *3 ("This gets discovery backward. Discovery is not a 'fishing expedition.'") (citation omitted).[4]

Plaintiffs Identify No Error Regarding NVIDIA's State Of Mind Concerning The Other "Pirated Libraries." Plaintiffs rehash the same argument the Magistrate Judge rejected—that whether NVIDIA discussed the other "pirated libraries" is somehow relevant to willfulness. ECF 162 at 3-4.[5] But the only relevant intent for willfulness under the Copyright Act is NVIDIA's state of mind as to Plaintiffs' asserted works—*i.e.*, works in the Books3 dataset. ECF 162 at 8. In addition, NVIDIA already agreed to run search terms related to piracy generally (guilt, steal, violat*, pirat*, pirac*), providing adequate discovery on Plaintiffs' willfulness theories. ECF 162-2 at 6-11. Plaintiffs cite no authority for the proposition that purported ***general*** knowledge or discussion of infringement by NVIDIA of third-parties outside the defined putative class is relevant to whether NVIDIA willfully infringed ***Plaintiffs' works***.

CONCLUSION

Plaintiffs ask the Court to order NVIDIA to run search terms and produce documents that relate exclusively to unpled claims. Plaintiffs have not met their burden to show that Magistrate Judge Kim's Order was clearly erroneous, let alone that discovery into LibGen, Anna's Archive, Sci-Hub, E-Libra, PiLiMi, or Z-Library (B-ok)—datasets untethered to their allegations—is proportional to the needs of the case. NVIDIA respectfully requests that the Court deny Plaintiffs' request to reverse the Magistrate Judge's Order and sanction a fishing expedition.

---

[4] *Rodas v. Monetary Mgmt. of Cal., Inc.*, was a denial of a motion to strike class allegations on the ground that it was premature. 2015 WL 1440602, at *4 (E.D. Cal. Mar. 27, 2015). It is inapposite because NVIDIA has not requested the same relief, nor did the Order strike any such allegations.

[5] Plaintiffs raise a new argument that discussions of using copyrighted books "is relevant to the market for books to train LLMs . . . under the fourth fair use factor" (Mot. at 5). Plaintiffs did not present that argument to the Magistrate Judge and cannot now rely on it to challenge the Order. ECF 162 at 1-4. *See Symantec Corp. v. Zscaler, Inc.*, 2019 WL 8331428, at *2 n.1 (N.D. Cal. Sept. 19, 2019) (declining to consider new argument not presented to the magistrate judge because the Magistrates Act was not "intended to give litigants an opportunity to run one version of their case past the magistrate, then another past the district court") (citation omitted). Plaintiffs also fail to explain how mere ***discussions*** of using copyrighted books could be relevant to the fourth factor.

Dated: September 19, 2025

Respectfully Submitted,

By: */s/ Andrew H. Schapiro*

QUINN EMANUEL URQUHART & SULLIVAN, LLP

Sean S. Pak (SBN 219032)
seanpak@quinnemanuel.com
50 California Street, 22nd Floor
San Francisco, CA 94111
Telephone: (415) 875-6600
Facsimile: (415) 875-6700

Andrew H. Schapiro (admitted *pro hac vice*)
andrewschapiro@quinnemanuel.com
191 N. Wacker Drive, Suite 2700
Chicago, Illinois 60606
Telephone: (312) 705-7400
Facsimile: (312) 705-4001

Alex Spiro (admitted *pro hac vice*)
alexspiro@quinnemanuel.com
295 Fifth Avenue
New York, NY 10016
Telephone: (212) 849-7000
Facsimile: (212) 849-7100

Rachael L. McCracken (SBN 252660)
rachaelmccracken@quinnemanuel.com
865 South Figueroa Street, 10th Floor
Los Angeles, CA 90017
Telephone: (213) 443-3000
Facsimile: (213) 443-3100

Attorneys for Defendant,
NVIDIA Corporation