May 15, 2025

The Honorable Lisa J. Cisneros
United States District Court for the Northern District of California
450 Golden Gate Ave.,
San Francisco, CA 94102

Re: *In Re Mosaic LLM Litigation*; Case No. 3:24-cv-01451-CRB

Dear Magistrate Judge Cisneros,

The parties submit this letter brief regarding a dispute pertaining to DBRX, a Databricks large language model ("LLM"). The parties met and conferred but were unable to reach a resolution on this issue. The relevant discovery dates and case management deadlines are below:

(1) Close of Fact Discovery on Plaintiffs' copyright claim/fair use: September 30, 2025.
(2) Close of Expert Discovery: January 23, 2026.
(3) Summary Judgment Motions on Plaintiffs' copyright claim/fair use: February 25, 2026.
(4) Oppositions to Summary Judgment Motions: April 7, 2026.
(5) Replies to Summary Judgment Motions: May 6, 2026.

Plaintiffs' Position

Defendants object on relevance grounds to any discovery relating to DBRX, Defendants' latest LLM released one month before the last complaint filed in this case. While DBRX is not referenced in the Complaints, Plaintiffs specifically allege that Defendants' infringing conduct is ongoing, including with the continued training of LLMs. *See* ECF No. 1 at ¶35. Defendants' position on relevance—a blanket objection to unilaterally decide what, if any, information about DBRX to produce—is inconsistent with Plaintiffs' allegations, the Federal Rules, and is extraordinarily inefficient. Defendants cannot insulate themselves from discovery by rebranding the quick developing technology at the heart of this dispute.

Relevance for discovery purposes "is broadly defined to encompass 'any matter that bears on, or that *reasonably could lead to* other matter that could bear on, any issue that is or *may be in the case*.'" *Doe v. Kaiser Found. Health Plan, Inc.*, No. 23-CV-02865, 2024 WL 3225904, at *1 (N.D. Cal. June 28, 2024) (citation omitted) (emphasis added). And "[r]elevancy should be construed liberally and with common sense . . . ." *Pivetti v. Mercedes Benz USA LLC*, No. 5:23-CV-02148-SB-SP, 2024 WL 943943, at *2 (C.D. Cal. Feb. 6, 2024) (cleaned up). Common sense leads to one conclusion: DBRX is relevant for purposes of discovery in this case.

Background of DBRX. The first complaint in this case was filed on March 8, 2024, alleging Defendants infringed the copyrights of at least tens of thousands of authors in the development of their MPT series LLMs. *See* ECF No. 1. The MPT series are "GPT-style decoder-only transformer[]" LLMs created by MosaicML and distributed by Databricks. *See* The Mosaic Research Team, *MPT-30B*, Databricks (June 22, 2023), https://www.databricks.com/blog/mpt-30b ("MPT-30B Press Release"). Defendants publicly identified the training data used for these LLMs—For example, the MPT-30B model included the "Redpajama - Books" dataset, which contains pirated material. *See id.*; *see also* ECF No. 1 at 5–6.

DBRX, released on March 27, 2024, is a "decoder-only large language model" created by that same MosaicML team and distributed by Databricks. *See* The Mosaic Research Team, *Introducing DBRX*, Databricks (Mar. 27, 2024), https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm ("DBRX Press Release"). For the DBRX models, Defendants *have not* publicly identified the training data. When the VP of AI at Databricks and former CEO of MosaicML was asked if DBRX used copyrighted training data, "he didn't answer directly" but stated only that they used "open data sets that the community knows." Kyle Wiggers, *Databricks Spent $10M on New DBRX Generative AI Model*, TechCrunch (Mar. 27, 2024 5:00AM).

DBRX is relevant. Defendants cannot avoid discovery simply because their newest model was released under a different name. The "allegations in a complaint generally dictate what evidence is discoverable . . . [but] the scope of permissible discovery is not based solely on whether a transaction is expressly mentioned in the complaint." *Scherer v. FCA US, LLC*, 538 F. Supp. 3d 1002, 1005 (S.D. Cal. 2021) (citations omitted) (holding defendants' "knowledge of Stalling Defects in Chrysler Pacifica vehicles [beyond the type of vehicle specifically at issue] is relevant"). Likewise, Plaintiffs alleged that since Databricks acquired MosaicML, MosaicML has continued to make copies of the Infringed Works for *LLM training* and other commercial purposes. ECF No. 1 at ¶35 (emphasis added). This allegation is not limited to the MPT models and therefore encompasses DBRX.

Discovery of DBRX is essential to understand the scope of Defendants' infringing activities. Defendants pirated hundreds of thousands of books to train the MPT models. Where Defendants could not obtain 1 trillion tokens for training the MPT-30b model without turning to pirated content—the same could be reasonably inferred for DBRX, which was trained on *12 trillion tokens*. DBRX Press Release*, supra*. A single token represents a word or character. *See* Dave Salvator, *Explaining Tokens*, Nvidia (Mar. 17, 2025). The DBRX models fit squarely into this litigation. In another case alleging that Meta Platforms, Inc.'s ("Meta") LLMs infringe copyright, the court overruled Meta's objection to producing discovery about a newly developed LLM not referenced in the complaint. *Kadrey v. Meta*, 3:23-cv-03417 (N.D. Cal.), ECF No. 267 at 11, ECF No. 279 at 4. And in *Capture Eleven, LLC v. Otter Products*, *LLC*, a court specifically granted discovery to identify additional instances of copyright infringement not specifically mentioned in the complaint. No. 20-CV-02551, 2021 WL 12298517, at *3–*4 (D. Colo. May 11, 2021) (collecting cases allowing discovery to identify infringement against products not specified in the complaint). That the court also denied discovery to the plaintiff's unasserted and unregistered copyrights, as Defendants point to, is simply irrelevant here. *See id.* at *5.

Defendants' attempt to prevent discovery about DBRX is unavailing. Defendants themselves stated that DBRX is the culmination of "years of LLM development at Databricks that includes the MPT . . . projects." DBRX Press release, *supra*. And Defendants have connected the training of DBRX to the MosaicML research team. *Id.* Thus, defendants' reliance on the ruling in *Authors Guild v. OpenAI*, No. 23-cv-8292 (S.D.N.Y. December 6, 2024), ECF No. 29, denying discovery into Microsoft's unalleged LLMs on the basis that they were entirely independent of OpenAI's accused models, is inapposite. And courts *in this district* have ordered OpenAI to produce discovery regarding new LLMs, instructing that "discovery into the recent and in-development [LLMs] appeared reasonable." *Tremblay v. OpenAI, Inc.*, No. 23-cv-03223, 2025 WL 84635, at *2 (N.D. Cal. Jan. 13, 2025). Here, Defendants are corporate affiliates and the direct

infringement claim and other allegations relate to MosaicML's infringing acts by training LLMs—Mosaic trained and built DBRX, too. Plaintiffs cannot be faulted where Defendants have refused to provide discovery on how DBRX was trained. Defendants' other cases are similarly unavailing. *See Micro Motion, Inc. v. Kane Steel Co.*, 894 F.2d 1318, 1328 (Fed.Cir.1990) (patent case); *Samsung SDI Co. v. Matsushita Elec. Indus. Co.*, No. CV 05-8493-AG, 2007 WL 4357552, at *5 (C.D. Cal. June 25, 2007) (noting *patent rules* "may narrow the scope of required discovery"); *Lineberry v. Addshoppers, Inc.*, No. 23-CV-01996-VC (PHK), 2024 WL 4707986, at *4 (N.D. Cal. Nov. 6, 2024) (noting the requested discovery related to "distinguishable activity" unrelated to the issues in the case). Defendants' ongoing infringement is not "distinguishable activity." *Lineberry*, 2024 WL 4707986, at *4.

Defendants improperly objected categorically to discovery on DBRX. Defendants objected to any search term that included DBRX. *See* ECF No. 119-2 & 3 (rows 28, 36–37). They later agreed to run the terms but stated they would not produce responsive documents. *See* ECF No. 119 at 5. When asked to produce training data for LLMs including DBRX, Defendants only agreed to provide training data for the MPT models. Ex. A at 4 (RFP No. 15). In an effort to avoid this dispute, Defendants amended select discovery responses to state that they did not use Books3 or "Redpajama-Books" datasets to train DBRX. *See, e.g.*, Ex. B at 3–6 (RFAs 6 & 10). That neither resolves the issue of relevance, nor undermines Plaintiffs' relevance arguments. These two datasets are only two of many relevant pirated databases at issue here. *See* ECF No. 1 at ¶28 (identifying numerous other sources of pirated content used by AI companies); *see also Kadrey*, ECF 574 at 10–11 (Plaintiffs describing evidence related to Meta's use of these same sources to train its LLMs). Defendants cannot avoid a determination of relevance by unilaterally and selectively allowing information to trickle out. *See Kadrey*, ECF No. 315 at 7 (Meta may "not decline to produce otherwise responsive documents on the ground that they concern" a new LLM under development). Lastly, Defendants have no burden argument at this time. They have already agreed to run DBRX-specific search terms, which generated minimal additional hits. For other search terms, Defendants are already reviewing those documents, and so requiring production of DBRX-related documents adds no burden. Plaintiffs cannot obtain information on the training or training data for DBRX elsewhere given Defendants' newfound secrecy.

Final Compromise: Plaintiffs respectfully ask the Court to order Defendants to produce responsive information, documents, and training data related to DBRX, or alternatively, order discovery as to DBRX's training data and sources of training data.

Defendants' Position

Plaintiffs seek to avoid their Rule 8 (and Rule 11) obligations by demanding discovery into DBRX models ("DBRX") that Plaintiffs do not mention, much less accuse of infringement, in their Complaints. DBRX originated entirely from a party (Databricks) that Plaintiffs *do not accuse of direct infringement*. The Court should reject Plaintiffs' improper attempt to obtain sweeping discovery on Databricks' models when Plaintiffs have not even attempted to plead a viable claim.

Background. This case is, and has always been, about *MosaicML's* alleged use of a specific dataset called "Books3" to train LLMs called the MPT models, which took place before Databricks acquired MosaicML in 2023. The consolidated action consists of two virtually

identical complaints, filed in March 2024 and May 2024. Dkt. 1; *Makkai v. Databricks*, No. 3:24-cv-02653 (N.D. Cal. May 2, 2024), Dkt. 1. The Complaints allege that the MPT models were trained on Books3 (itself part of a larger dataset called "RedPajama – Books") that allegedly includes Plaintiffs' books. Dkt. 1 ¶¶ 24-35. Based on those factual allegations, Plaintiffs brought a direct infringement claim against MosaicML over its training of the MPT models. *Id.* ¶¶ 36-42. Importantly, Plaintiffs' sole claim against Databricks is *not* for direct infringement, but for *vicarious* liability—stating that as the corporate parent, Databricks had the "right and ability to control" and benefitted financially from *MosaicML's* alleged infringements. *Id.* ¶¶ 43-46.

Now, a year after filing the Complaints, Plaintiffs seek sweeping discovery about DBRX. Far from a mere relabeling of technology, Databricks (not MosaicML) trained and released DBRX using a different mix of data than MosaicML's MPT models. Plaintiffs' Complaints contain *no allegations* about DBRX, even though their second complaint was filed *after* DBRX's release. And in the year since then, Plaintiffs made no attempt to amend to add claims about DBRX or any direct infringement claim against Databricks. Nonetheless, to try to resolve this dispute, Defendants amended their verified interrogatory responses and answered RFAs to make clear that DBRX was *not* trained on Books3. Yet Plaintiffs *still* seek full discovery on DBRX.

Plaintiffs' DBRX discovery requests are irrelevant to their allegations and claims. It is well settled that a plaintiff must first assert a plausible claim supported by factual allegations that satisfies Rule 8 *before* seeking discovery to support it. If Plaintiffs had a basis to bring direct infringement claims against Databricks over DBRX, they should have filed complaints asserting those claims. Because they have not done so—and on the facts here, cannot do so—the Court should not permit discovery into DBRX.

To satisfy Rule 26, the discovery sought must relate to a party's claim or defense. *In re Williams-Sonoma, Inc.*, 947 F.3d 535, 539 (9th Cir. 2020) (explaining how Rule 26 was amended to restrict discovery and *eliminated* the "subject matter" reference). This makes sense, as "the discovery rules are designed to assist a party to prove a claim it reasonably believes to be viable *without discovery*, not to find out if it has any basis for a claim." *In re Countrywide Fin. Corp. Mortg.-Backed Sec. Litig.*, No. 2:11-CV-10549 MRP, 2013 WL 5614294, at *8 (C.D. Cal. Sept. 30, 2013) (quoting *Micro Motion, Inc. v. Kane Steel Co.*, 894 F.2d 1318, 1327 (Fed.Cir.1990)).

Courts have repeatedly held that parties have no right to discovery on items not in their pleadings. *See, e.g.*, *Lineberry v. Addshoppers, Inc.*, No. 23-CV-01996-VC (PHK), 2024 WL 4707986, at *3 (N.D. Cal. Nov. 6, 2024) (denying motion to compel). And alleging claims about one product does not make discovery about a separate product "relevant." *See, e.g.*, *Lin v. Solta Med., Inc.*, No. 21-CV-05062-PJH, 2023 WL 8374740, at *2 (N.D. Cal. Dec. 4, 2023) (limiting discovery to "materials *concerning the model or generation of device at issue in the complaint*—not any device with the [accused product's] name on it" (emphasis added)).

Indeed, in another litigation involving LLM training, Plaintiffs' counsel *twice* tried to pursue discovery about models not in their complaint (using largely the same arguments they make here), and the court *denied* both motions. *See Authors Guild v. OpenAI, Inc.*, No. 23-cv-8292 (S.D.N.Y. Dec. 6, 2024), Dkt. 293 at 2-3 (denying Plaintiffs' motions at Dkts. 270, 271). The court denied plaintiffs' first attempt to compel discovery about additional OpenAI models despite plaintiffs' blanket assertion that "all OpenAI models infringe." Dkt. 270 at 2. Plaintiffs also sought documents showing datasets that *Microsoft*, an investor in and partner of OpenAI, used to

train its LLMs when the allegations were against *OpenAI*. Dkt. 271. The court rejected both attempts. Dkt. 293 at 2-3; *see also* Dkt. 279 (Microsoft's response); Dkt. 281 (OpenAI's response).

The same outcome should occur here. Plaintiffs identify no allegations that would bring DBRX into the case. Instead, they recast their allegations as against "Defendants" when in fact they are against *MosaicML*. *See, e.g.*, Dkt. 1 ¶ 35. The blogpost that Plaintiffs cite made clear that DBRX is separate from the MPT models and distinguished DBRX's training data. That both models are "decoder transformer" models is also a red herring, as that technology is ubiquitous in AI research. And the fact that former MosaicML scientists worked on DBRX does not change that DBRX is a separate model not trained on Books3, the sole dataset on which Plaintiffs' infringement allegations are based. Further, Plaintiffs' cited cases provide them no support. *Kadrey* is inapplicable as there was no dispute; Meta told the court that it was not withholding documents related to an unreleased LLM. *Kadrey* Dkts. 267 at 11, Dkt. 279 at 4. In *Tremblay*, the court adopted *Defendants'* compromise proposal for GPT-class models. 2025 WL 84635, at *2. *Scherer* allowed discovery about knowledge of issues *alleged in the complaint*. 538 F. Supp. at 1005. And *Capture Eleven* actually *supports* denying Plaintiffs' request, as that court held that discovery about images *not* in the complaint was "not proportional to needs of the case." 2021 WL 12298517, at *4.

Plaintiffs' direct infringement claim against MosaicML over training the MPT models with Books3 does not entitle them to scorched earth discovery about a *separate model released by a post-acquisition parent company that is not accused of infringing conduct*. Plaintiffs cannot identify anything in their Complaints to justify this obvious fishing expedition. At bottom, Plaintiffs argument boils down to conjecture that, because the MPT models were allegedly trained on Books3, DBRX must have been too. Tellingly, Plaintiffs have not sought to amend their Complaints on that flimsy theory, which would not satisfy Rules 8 or 11.

Plaintiffs' demand for all discovery about DBRX is disproportionate. Defendants have already provided verified discovery responses stating that DBRX was *not* trained on RedPajama – Books or Books3. And while Plaintiffs *now* assert that their case is about additional potential sources of their books, *that is not what their Complaints allege*, and Plaintiffs cannot constructively amend their pleadings through briefing. *See Hodge v. Travel + Leisure Co.*, No. 5:24-cv-06116-EJD, 2025 WL 327741, at *2 n.1 (N.D. Cal. Jan. 29, 2025) (briefing cannot amend a complaint); Dkt. 1 ¶¶ 25-42 (alleging infringement based only on "copy[ing] of the Books3 dataset"); *id.* ¶ 28 (mentioning "shadow libraries" *without alleging that MosaicML used them*).

Rule 26's proportionality requirement "is intended to encourage judges to be more aggressive in identifying and discouraging discovery overuse." *Lineberry*, 2024 WL 4707986, at *2 (quoting advisory committee notes). Defendants have already searched for, collected, and produced more than 40 *terabytes* of data—the equivalent of 10 *billion* pages (enough to circle the earth more than 50 times)—based on Plaintiffs' requests concerning the MPT models. And as Plaintiffs acknowledge, DBRX was trained on far more data. Requiring Defendants to replicate this costly exercise for an entirely separate model, untethered to any allegation in their Complaints, is exactly the type of disproportionate discovery that the Rules preclude.

Final Compromise: To resolve this dispute, Defendants provided discovery responses showing that Books3 was not used to train DBRX. Defendants have asked Plaintiffs if they would agree to any limitation on DBRX discovery, and Plaintiffs have refused to offer any compromise. The Court should deny Plaintiffs' requests for discovery on DBRX.

DATED: May 15, 2025

By: */s/ Joseph R. Saveri*

Joseph R. Saveri (SBN 130064)
Christopher K.L. Young (SBN 318371)
Evan Creutz (SBN 349728)
Elissa A. Buchanan (SBN 249996)
William Castillo Guardado (SBN 294159
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
cyoung@saverilawfirm.com
ecreutz@saverilawfirm.com
eabuchanan@saverilawfirm.com
wcastillo@saverilawfirm.com

Matthew Butterick (SBN 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
Telephone: (323) 968-2632
Facsimile: (415) 395-9940
mb@buttericklaw.com

Justin A. Nelson (admitted *pro hac vice*)
Alejandra C. Salinas (admitted *pro hac v*
SUSMAN GODFREY L.L.P
1000 Louisiana Street, Suite 5100
Houston, TX 77002-5096
Telephone: (713) 651-9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (SBN 316062)
SUSMAN GODFREY L.L.P
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067-2906
Telephone: (310) 789-3100
RNath@susmangodfrey.com

Elisha Barron (admitted *pro hac vice*)
Craig Smyser (admitted *pro hac vice*)
SUSMAN GODFREY L.L.P
One Manhattan West, 51st Floor

By: */s/ Jedediah Wakefield*

Jedediah Wakefield (CSB No. 178058)
jwakefield@fenwick.com
Ryan Kwock (CSB No. 336414)
rkwock@fenwick.com
FENWICK & WEST LLP
555 California Street, 12th Floor
San Francisco, CA 94104
Telephone: 415.875.2300
Facsimile: 415.281.1350

David Hayes (CSB No. 122894)
dhayes@fenwick.com
FENWICK & WEST LLP
801 California Street
Mountain View, CA 94041
Telephone: 650.988.8500
Facsimile: 650.938.5200

Deena Feit (admitted *pro hac vice*)
dfeit@fenwick.com
FENWICK & WEST LLP
401 Union Street, 5th Floor
Seattle, WA 98101
Telephone: 206.389.4510
Facsimile: 206.389.4511

Charles Moulins (admitted *pro hac vice*)
cmoulins@fenwick.com
FENWICK & WEST LLP
902 Broadway, Ste 14
New York, NY 10010
Telephone: 212.430.2600
Facsimile: 650.938.5200

Zachary Harned (CSB No. 335898)
zharned@fenwick.com
FENWICK & WEST LLP
730 Arizona Avenue, 1st Floor
Santa Monica, CA 90401
Telephone: 310.554.5400
Facsimile: 650.938.5200

New York, NY 10019
Telephone: (212) 336-8330
ebarron@susmangodfrey.com
csmyser@susmangodfrey.com

Attorneys for Defendants
DATABRICKS, INC., and
MOSAIC ML, LLC, formerly
MOSAIC ML, INC.

Jordan W. Connors (pro hac vice)
Trevor D. Nystrom (pro hac vice)
SUSMAN GODFREY L.L.P
401 Union Street, Suite 3000
Seattle, WA 98101
Telephone: (206) 516-3880
jconnors@susmangodfrey.com
tnystrom@susmangodfrey.com

Rachel J. Geman (admitted *pro hac vice*)
Danna Z. Elmasry (admitted *pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, NY 10013
Tel.: 212.355.9500
rgeman@lchb.com
delmasry@lchb.com

Anne B. Shaver
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111
Tel.: 415.956.1000
ashaver@lchb.com

Betsy A. Sugar (admitted *pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
222 2nd Avenue S. Suite 1640
Nashville, TN 37201
Tel.: 615.313.9000
bsugar@lchb.com

Rachel J. Geman (admitted *pro hac vice*)
Danna Z. Elmasry (admitted *pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, NY 10013
Tel.: 212.355.9500
rgeman@lchb.com
delmasry@lchb.com

Bryan L. Clobes (admitted *pro hac vice*)
Alexander J. Sweatman (admitted *pro hac vice*)
Mohammed A. Rathur (admitted *pro hac vice*)
CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
135 South LaSalle Street, Suite 3210
Chicago, IL 60603
Telephone: 312-782-4880
bclobes@caffertyclobes.com
asweatman@caffertyclobes.com
mrathur@caffertyclobes.com

Brian D. Clark (admitted *pro hac vice*)
Laura M. Matson (admitted *pro hac vice*)
Arielle Wagner (admitted *pro hac vice*)
LOCKRIDGE GRINDAL NAUEN PLLP
100 Washington Avenue South, Suite 2200
Minneapolis, MN 55401
Telephone: (612) 339-6900
Facsimile: (612) 339-0981
Email: bdclark@locklaw.com
lmmatson@locklaw.com
aswagner@locklaw.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class

ATTESTATION OF CONCURRENCE IN FILING PURSUANT TO CIVIL L.R. 5-1(i)(3)

This document is being filed through the Electronic Case Filing (ECF) system by attorney Rohit Nath. By their signature, Rohit Nath attests that he has obtained concurrence in the filing of this document from each of the attorneys identified in the above signature block.

Dated: May 15, 2025

By: */s/ Rohit D. Nath*
Rohit D. Nath