DAVID H. KRAMER, SBN 168452
Email: dkramer@wsgr.com
MAURA L. REES, SBN 191698
Email: mrees@wsgr.com
WILSON SONSINI GOODRICH & ROSATI
Professional Corporation
650 Page Mill Road
Palo Alto, CA 94304-1050
Telephone: (650) 493-9300

ERIC P. TUTTLE, SBN 248440
Email: eric.tuttle@wsgr.com
WILSON SONSINI GOODRICH & ROSATI
Professional Corporation
701 Fifth Avenue, Suite 5100
Seattle, WA 98104-7036
Telephone: (206) 883-2500

PAUL J. SAMPSON, admitted *pro hac vice*
Email: psampson@wsgr.com
WILSON SONSINI GOODRICH & ROSATI
Professional Corporation
95 S. State Street
Suite 1000
Salt Lake City, Utah 84111
Telephone: (801) 401-8510

Counsel for Defendant
GOOGLE LLC

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

In re Google Generative AI Copyright Litigation	Master File Case No.: 5:23-cv-03440-EKL Consolidated with Case No.: 5:24-cv-02531-EKL **DEFENDANT GOOGLE LLC'S OPPOSITION TO PLAINTIFFS' MOTION TO AMEND CASE SCHEDULE** Judge: Hon. Eumi K. Lee Magistrate Judge: Susan van Keulen

TABLE OF CONTENTS

I. INTRODUCTION

The Court has already seen this motion. The last time Plaintiffs were days away from the class certification deadline, they pointed fingers, mischaracterized the record, and pleaded for months more time while they continued hunting for some plausible justification. *See* ECF No. 173. In response, this Court gave Plaintiffs six extra weeks, while admonishing them for "wast[ing] time on distractions," "seeking overbroad discovery, [and] failing to prioritize issues relevant to class certification." ECF No. 197 at 2. Yet here we are again, days away from the new deadline, with Plaintiffs spinning out the same false narrative and seeking another lengthy extension.

The deja vu is particularly powerful because Plaintiffs presented the very same supposed grievances supporting this extension request to Judge van Keulen just last week, seeking a laundry list of relief, including appointment of a special master. They were rebuffed, with Judge van Keulen scolding them twice in a brief order. ECF No. 221. As Judge van Keulen recognized, what Plaintiffs imply were some sort of obstructions with respect to training data are merely the inevitable hiccups inherent in reviewing information at the virtually unprecedented scale Plaintiffs demanded and received. The only concrete issue Plaintiffs identified to Judge van Keulen and again here—a format conversion error affecting roughly only three percent of the training data Plaintiffs requested and Google provided—was proactively discovered, disclosed, and corrected by Google within days. Moreover, the data Plaintiffs falsely claim was "withheld" due to the error was actually produced and accessible to them all along in the original format. Beyond that, the record demonstrates that Plaintiffs' claimed lack of progress is a function of self-inflicted technical problems and continuing lack of diligence. Google has provided ample assistance and made its technical advisors available as directed.

Regardless, Plaintiffs' complaints about training data have no bearing on class certification. Plaintiffs assert that their analysis of training data is essential to determining whether their works and others were used. But that is not genuinely contested at this stage. Indeed, at the Court's urging, to simplify the issues, Google has stipulated that it will not dispute for purposes of class certification that Plaintiffs' works were present in Google's training data. ECF No. 222. Plaintiffs' notion that they could or should now be trying to catalog every work in the datasets (as opposed to arguing they

could feasibly do so later based on a sample) is both wrong and unnecessary. If that exercise ever matters, it would be at the merits stage. At this point, a supposed identification of works in the training data is no answer for the individualized issues inherent in Plaintiffs' claims that overwhelmingly predominate and should preclude certification.

Plaintiffs' motion for yet another extension should be denied. If the Court is inclined to grant any relief, it should adopt Google's good-faith compromise offer of a brief, three-week extension limited to class certification deadlines only, with no downstream schedule adjustments.

II. BACKGROUND

Plaintiffs' "Relevant Background" reprises, often verbatim, the same mischaracterizations about Google's training-data production from their Motion to Enforce Discovery Order and to Appoint a Special Master (ECF No. 208). Google already addressed those assertions (ECF No. 214). At the September 11 hearing on that motion and in her September 12 Order, Judge van Keulen largely denied Plaintiffs' requested relief, ordering only narrow, practical assistance (*see* ECF No. 227 ("Sept. 11, 2025 Hr'g Tr.") 44:14-46:2; ECF No. 221). Since that hearing, Google's engineers met twice with Plaintiffs' experts and showed why many of Plaintiffs' supposed "problems" were actually the product of their own errors or misunderstandings, including a failure to follow simple instructions. Google briefly addresses several of these points again here.

A. Plaintiffs' Critiques of the Review Environment Are Unsupported.

At a June 18 discovery hearing, Judge van Keulen directed Plaintiffs' inspection of training data "should begin as early as next week." June 18, 2025 Hr'g Tr. 34:4-5. Google met that timeline. After Plaintiffs provided mailing addresses for reviewers on June 20, Google shipped them Chromebooks; after Plaintiffs provided email addresses on June 24, Google created accounts and credentials for them. By Friday, June 27, the review environment that Google is hosting at its own considerable expense was accessible to Plaintiffs. Plaintiffs confirmed receipt and that they had logged in on June 28 and requested additional instructions, which Google provided the next business day. When Plaintiffs reported on July 1 that some of their users could not access the environment, Google sent further guidance on July 2 that resolved the issue. *See* Decl. of Paul J. Sampson ("Sampson Decl.") ¶ 3.

Plaintiffs identified 23 training datasets for hosted review on Google's servers. On June 25, Google advised that four of those datasets were not linked to any model at issue and committed to produce the remaining 19. Given the datasets' enormous size and distributed storage, locating and copying them into the review environment required some time. By June 27, 12 of the 19 datasets were live; by July 10, all 19—comprising about 1.5 petabytes of data—were available to Plaintiffs and have remained available ever since. *Id.* ¶ 2.

Thereafter, Google provided extensive support for Plaintiffs' access. On June 20, Plaintiffs asked that more than 40 software libraries be installed in the review environment. Google promptly installed the vast majority, declining only a few that failed its security or compatibility review. To further ease Plaintiffs' analysis, Google also installed format-specific libraries for two of the formats in which the data was stored without even being asked. *Id.* ¶ 4.

Plaintiffs requested a call with Google technical personnel on July 8; Google agreed on July 9 and asked for topics so the appropriate engineers could attend. Plaintiffs claim that "Google did not make its technical personnel available to confer with Plaintiffs' experts until July 29, 2025." Mot. at 4. But they neglect to say that they did not provide topics for inquiry so Google could provide the right people until July 21. Google supplied informal written responses to Plaintiffs' inquiries on July 22, and the parties held what Plaintiffs described as a "productive" call on July 29. The next day, to assist Plaintiffs' experts, Google provided example code that Google's engineers had shown Plaintiffs' experts during the call. At Plaintiffs' request, a second technical session was held on August 5. Google has continued to provide written guidance to Plaintiffs and promptly process requests to reset their access to devices. Sampson Decl. ¶ 6.

Plaintiffs complain about intermittent "outages." Mot. at 7, 9. But the datasets they chose total ***more than a petabyte*** of data, and no secure cloud environment can support broad, continuous code-based queries at that scale without occasional slowdowns or resets. Google explained these constraints at the outset; Plaintiffs nevertheless chose to load nearly 20 enormous datasets simultaneously. When interruptions occurred, Google restored access promptly. Plaintiffs, by contrast, declined to identify the queries running at the time of each freeze, preventing diagnosis of whether the interruptions stemmed from their own resource-intensive processes (which include jobs

that run through "the middle of the night" or lasting "approximately 23 hours," ECF No. 223-2, McCarron Decl. ¶ 11). Resets must be performed by Google personnel during business hours, yet Plaintiffs often reported problems late on Friday evenings or over holiday weekends and then faulted Google for not resolving them immediately.[1] Sampson Decl. ¶¶ 7-8. As Judge van Keulen observed from her experience, however, large-dataset discovery "always" has "wrinkles," June 18, 2025 Hr'g Tr. at 63:10-12, and "there may not be a comprehensive fix because … [i]t's not a bug … [i]t's a feature … [of] managing these massive amounts of data," Sept. 11, 2025 Hr'g Tr. at 7:18-8:6.

Several of Plaintiffs' specific complaints suggest unfamiliarity with the technical requirements of petabyte-scale analysis. For example, Plaintiffs contend that Google "declined" to provide them with "BigQuery" and "insisted" on "TensorFlow." Mot. at 4. BigQuery is a warehousing/query service; TensorFlow is a machine learning software library used to read and process files (including TFRecords). Plaintiffs are simply wrong about BigQuery; they have always had write access to BigQuery if they wish to use it to analyze the data. Equally troubling is Plaintiffs confusing the concept of a storage/query platform with a file-format parsing tool. Plaintiffs' confusion is not an "access" problem; it reflects that Plaintiffs do not appreciate basic mechanics of the review process. Sampson Decl. ¶ 5. Another example appears in Plaintiffs' account of system "outages." Mot. at 7, 9. In their motion before Judge van Keulen, Plaintiffs' expert Ayyub Ibrahim reported a "continuing outage" beginning September 5 (ECF No. 208-3 ¶ 4); in support of ***this*** motion, he says the "continuing outage … was only just resolved today [September 12] … after a technical call" with a Google engineer (ECF No. 223-3 ¶ 4). What he omits is that the Google engineer "resolved" the issue by walking Ibrahim through a standard Chromebook "powerwash," consistent with written instructions Google provided weeks earlier that are available online;[2] had Mr. Ibrahim followed those steps when provided, the "outage" would have been avoided altogether. *See* Sampson Decl. ¶ 18.

[1] As discussed at the September 11 hearing, Google is developing a self-service reset tool for the environment. *See* Sept. 11, 2025 Hr'g Tr. 9:23-10:10. Rollout is targeted for early this week, pending validation testing; until then, resets must be performed by Google personnel during business hours. Sampson Decl. ¶ 8.

[2] *See* Chromebook Help page, *Reset your Chromebook to factory settings*, available at https://support.google.com/chromebook/answer/183084?hl=en.

B. Plaintiffs Mischaracterize a Narrow Conversion Issue that Google Identified and Corrected Within Days.

Plaintiffs asked that the web-crawled training datasets that Google supplied include URL metadata identifying the locations at which Google found the content. Google agreed to provide URLs where they existed in the underlying data, and most of the web-crawled datasets that Google provided contained them. That gave Plaintiffs hundreds of millions of content-URL pairs to analyze for whatever representative purpose they believe relevant to class certification. Four web-crawled training datasets selected by Plaintiffs (out of the 19 datasets Google provided) did not include URL metadata. For those four, Google went beyond Plaintiffs' requests by looking for upstream datasets from which the training datasets were derived. Google was able to locate these precursor sets that included URLs for three of the four datasets and uploaded those to the environment (the last finishing August 1). No precursor with URLs could be located for the fourth. Sampson Decl. ¶ 9.

Some of the datasets Plaintiffs requested resided in a format called "SSTable." While Plaintiffs could access those datasets and the URLs they contained, to make them more accessible and manipulable, Google proactively—without any request from Plaintiffs—also converted those datasets into a second format (TFRecord). Google uploaded the converted files in separate folders alongside the original SSTable datasets. Google began this effort immediately; more than half of the files were converted by July 10, with the remainder completed around August 7. *Id.* ¶ 10.

Plaintiffs' primary complaint concerns this voluntary conversion. *See* Mot. at 5-6. The original SSTable format stores data in a table-like format. Each record is a row, and the actual data for that record is stored in a column called the "value." There is also a second column called the "key," containing an alphanumeric sequence as an identifier for each row. The key is used to sort the rows into alphanumeric order, allowing very large datasets to be accessed efficiently. The Google engineer who converted the SSTables into TFRecords understood that the training data was contained in the "value" column, and copied only the "value" into the converted files. However, unbeknownst to her, the team that created one of the 19 datasets loaded to the review environment—called the "GCC" dataset—had included URL metadata in the key. *Id.* ¶ 11.

On August 19, a different Google engineer reported that one dataset converted from SSTable

to TFRecord was missing the alphanumeric identifier field. On Friday, August 22, Google's counsel learned from the engineer that the key field had been omitted during conversion, and asked Google engineers to investigate what information was contained in that field. On August 25-26, counsel for Google learned that the field in some cases contained URLs and immediately requested that all of the SSTables be converted again with the "key" field included. After exploring how to accomplish this quickly and accurately, Google engineers wrote the necessary code and began the process, adding newly converted files to the review environment in batches. Google notified Plaintiffs of the problem and solution on August 30, by which time all of the files with URLs in the key field had been reconverted, and finished reconverting the remaining files on September 2. *Id.* ¶ 12.

Plaintiffs inaccurately recounted this history to Judge van Keulen and misstated both its scope and effect. *First*, the original SSTable versions of all relevant datasets have been available to Plaintiffs since June or July, and ***any URLs in them have always been readable and accessible as plain text***. *Second*, of the 19 staged datasets, only one, the "GCC" set, stored URLs in the SSTable "key" field; accordingly, only GCC's TFRecord copies were affected—other datasets were either not in SSTable format or did not have URLs in the keys that were omitted during conversion. *Third*, Plaintiffs did not begin reviewing many of these files until late August anyway. On August 13, they claimed "31 datasets" were "corrupted or inaccessible." Mot. at 5. After Plaintiffs sent additional detail, Google determined the issue was Plaintiffs' mistaken assumption about the converted files' "string" structure—not the missing key field—and on August 21 Google explained the correct structure and directed Plaintiffs to example code that had been shared on July 30 that would have enabled parsing had Plaintiffs' experts used it. The "31 datasets" were in fact subsets or files within the GCC dataset—the only dataset affected by the conversion issue. Because reconversion to restore the key field was completed by August 30, even crediting Plaintiffs' account, the maximum impact was a nine-day delay confined to a subset of files within a single dataset, one of 19 total, and roughly three percent of the staged training data by file size. Sept. 11, 2025 Hr'g Tr. 30:7-31:15.

III. ARGUMENT

Plaintiffs have not shown "good cause" under Rule 16(b)(4) to extend the schedule yet again and they have not acted with diligence. Their complaints about the review environment are

overstated, and the only error they identify—a conversion issue that affected roughly three percent of the data—was promptly corrected. More fundamentally, the training-data disputes are tangential at best to the Rule 23 inquiry that Plaintiffs seem desperate to postpone. Indeed, Plaintiffs have not offered any persuasive (or even non-conclusory) explanation for how still more time with the data they have had for months will materially advance their class certification position.

A. Plaintiffs Cannot Show Good Cause or Diligence as Rule 16 Requires.

Plaintiffs' only effort to demonstrate good cause for another extension rests on a significant misrepresentation of the record. What they highlight as the longest "continuing outage" of access to training data (affecting Mr. Ibrahim) was resolved when a Google engineer walked him through a standard Chromebook powerwash—using steps Google had provided Plaintiffs weeks earlier (that Ibrahim had not previously followed). Plaintiffs also impeded troubleshooting efforts by refusing to identify the queries or code running during reported freezes, preventing diagnosis of whether interruptions stemmed from their resource-intensive jobs (which they acknowledge include 23-hour runs). Just last week Judge van Keulen agreed that such freezes and reboots are inherent to analysis of petabytes of data and are "not a bug … [but] a feature." Sept. 11, 2025 Hr'g Tr. at 7:18-8:17.

Plaintiffs' rhetoric about supposedly "critical [URL] metadata," Mot. at 4, tells the same story. Google's file conversion issue did not affect "49 datasets." *Id.* at 5. As Google explained in its briefing (ECF No. 214 at 4-5), at the September 11 hearing (Sept. 11, 2025 Hr'g Tr. at 30:7-32:5), and again on a follow-up call between a Google engineer and Plaintiffs' experts (Sampson Decl. ¶ 18), the "49 datasets" are actually files composing just a handful of the 19 datasets that Plaintiffs requested. Only 31 of those files—all belonging to ***a single dataset***, the "GCC" dataset—contained URLs in the key field that were not also present in the converted value field. By file size, the material affected by the error represents approximately ***three percent*** of the total training data provided to Plaintiffs. And most importantly, the URL data in the key field was not "withheld" at all. The original SSTables—with URLs readable as plain text—were available to Plaintiffs at all times. Plaintiffs cannot bootstrap their mischaracterization of this issue into "good cause."

Relatedly, Rule 16 requires a party seeking an extension to show diligence. *Johnson v. Mammoth Recreations, Inc.*, 975 F.2d 604, 609 (9th Cir. 1992). Here too, Plaintiffs fail. The last

time Plaintiffs sought an extension, the Court noted several discovery detours through which Plaintiffs had burdened Google and the Court, and cautioned them not to "waste[] time on distractions." ECF No. 197 at 2. Plaintiffs did not heed that directive and pursued additional discovery and motions that the Court rejected or narrowed substantially—*e.g.*, on RFAs, ECF No. 221 at 2 (rejecting Plaintiffs' maximalist positions on *hundreds* of RFAs and adopting Google's proposed compromises); another "overbroad, vague and ambiguous" Rule 30(b)(6) topic, *id.* at 3; and a request for a special master that Judge van Keulen declined before even seeing Google's response, ECF No. 211. That is not diligence. Granting still more time would encourage Plaintiffs to continue "engaging in discovery skirmishes on ancillary issues." ECF No. 197 at 3.

There is also the matter of the timing of this motion. On September 3, Plaintiffs announced they would file another motion for a schedule extension. They did not file it that day. Instead, they filed a motion on September 8 with Judge van Keulen. In it, they said they "ha[d] no choice but to seek an extension of the class certification deadline" and promised to file that motion "tomorrow, on September 9, 2025." ECF No. 208 at 3 n.1. But they did not file it that day either. Instead, they waited until just before midnight on Friday, September 12 to file—while insisting Google's opposition be due on Monday, September 15. If Plaintiffs truly "ha[d] no choice," they would have filed when they said they would. Their delay—paired with a late-night filing and an effort to compress Google's response into a weekend—suggests gamesmanship, not diligence.[3] Plaintiffs plainly waited to see if the September 11 hearing would produce soundbites supporting their desired extension. It did not; Judge van Keulen largely denied their requested relief and recognized their technical glitches as "the nature of [] the beast." Sept. 11, 2025 Hr'g Tr. at 44:14-19.

Having mischaracterized the record, continued to pursue meritless discovery motions, and delayed in seeking a further extension, Plaintiffs have not shown the diligence Rule 16 requires. Judge van Keulen presaged that Plaintiffs should not "count on getting more time." Sept. 11, 2025 Hr'g Tr. at 34:22. This repackaged request is an improper second bite at the same apple.

---

[3] The timing of the motion in the context of case discovery further underscores its tactical aim. The parties already have five depositions scheduled this week across multiple time zones. Plaintiffs' attempt to force an accelerated hearing on their extension motion, while discovery is in full swing, appears calculated to impose maximum burden on the Court and parties alike.

B. Plaintiffs Dramatically Overstate the Relevance of Training Data to Class Certification.

Plaintiffs contend they need six more weeks because "complete training data" is "indispensable" to class certification for two purposes: (1) to show that copyrighted works appear in Google's training data, and (2) to identify which specific copyrighted works were used. Mot. at 3. Neither contention warrants additional delay.

First, whether particular copyrighted works appear in the training data is not genuinely disputed for present purposes. Plaintiffs have long had extensive company documents, deposition testimony, and access to vast datasets (far beyond the narrow slice implicated by the conversion issue). Yet they have never explained why training data is necessary for class certification. The conclusory assertion in Plaintiffs' motion that their experts need to "execut[e] the analysts [sic] that will definitively show that [sic] Google's unfair infringement" goes to the merits rather than class certification (to the extent it is intelligible at all). Mot. at 8.

Second, Plaintiffs' demand for "direct evidence" from which they could try to identify every copyrighted work used, Mot. at 3, is precisely the sort of misguided and monumental burden that Google cited at the CMC last December, prompting the Court to repeatedly direct the parties to "really focus on discovery in terms of what is needed for the class cert to begin with." ECF No. 93 at 21:19-21. Plaintiffs have no need to identify every work Google used in the training process for class certification, and make no effort to explain why they would. Rule 23 asks whether common issues will predominate; it does not require a complete work-by-work inventory for "millions" of putative class members. Plaintiffs tacitly recognized as much by agreeing to receive only exemplary training datasets (albeit very large ones) for class certification discovery rather than all datasets for all models. Representative material is the appropriate tool at this stage. *See, e.g.*, *Valentine v. Crocs, Inc.*, 2024 WL 1636716, at *2 (N.D. Cal. Apr. 15, 2024) (pre-certification discovery should be limited to certification issues and avoid undue burden).

Regardless, more time with training data will not cure the structural problems in Plaintiffs' certification effort. Even if Plaintiffs could readily identify all works ever used for training, they would still have no way to deal with the myriad individualized proof problems raised by their claim

of infringement, including questions as to each work's ownership, chain of title, registration (validity and timing), licenses or consents (direct and indirect licenses, express and implied licenses, and licenses granted by agents or co-owners), statutes of limitation, fair use, damages, and failure to mitigate, which would overwhelm any common questions. *See Tyson Foods, Inc. v. Bouaphakeo*, 577 U.S. 442, 453 (2016) ("An individual question is one where members of a proposed class will need to present evidence that varies from member to member, while a common question is one where the same evidence will suffice for each member to make a prima facie showing [or] the issue is susceptible to generalized, class-wide proof."). From the start, Google has highlighted these issues, noting that courts cite them in recognizing that copyright claims are poor candidates for class treatment. *Schneider v. YouTube, LLC*, 674 F. Supp. 3d 704, 717 (N.D. Cal. 2023) ("Every copyright claim turns upon facts which are particular to that single claim of infringement …. Every copyright claim is also subject to defenses that require their own individualized inquiries."). Plaintiffs have never answered. Granting them another extension to rummage further through training data would reward their delay without advancing the certification analysis.

C. The Class Certification Schedule Should Hold

At some point, and Google submits that point is now, the class certification deadline must be given effect. Plaintiffs have already received a six-week extension (*see* ECF No. 197), and further extensions will only invite more motion practice, along with demands from Plaintiffs for more time when the next filing deadline approaches.

Although Google believes no extension is warranted, it offered a three-week compromise, mindful of this Court's guidance that it is "not keen to issue a middle-ground ruling on every dispute" and incentivize extreme positions. ECF No. 197 at 3. But Plaintiffs refused. Sampson Decl. ¶ 16. At this point, for the reasons above, their motion for a further extension should be denied. If the Court is inclined to grant relief, it should adopt Google's three-week compromise and limit an extension to class-certification deadlines only, with no other schedule adjustments and no further extensions absent a concrete, non-self-inflicted showing of good cause. Given that five depositions are proceeding this week across multiple time zones, *id.* ¶ 19, Plaintiffs' request for an expedited hearing should be denied; the motion can be decided on the papers or set on a regular schedule.

Respectfully submitted,

Dated: September 16, 2025

WILSON SONSINI GOODRICH & ROSATI
Professional Corporation

By: */s/ Paul J. Sampson*

Paul J. Sampson, admitted *pro hac vice*
Email: psampson@wsgr.com
WILSON SONSINI GOODRICH & ROSATI
Professional Corporation
95 S. State Street
Suite 1000
Salt Lake City, Utah 84111
Telephone: (801) 401-8510

Eric P. Tuttle, SBN 248440
Email: eric.tuttle@wsgr.com
WILSON SONSINI GOODRICH & ROSATI
Professional Corporation
701 Fifth Avenue, Suite 5100
Seattle, WA 98104-7036
Telephone: (206) 883-2500

David H. Kramer, SBN 168452
Email: dkramer@wsgr.com
Maura L. Rees, SBN 191698
Email: mrees@wsgr.com
WILSON SONSINI GOODRICH & ROSATI
Professional Corporation
650 Page Mill Road
Palo Alto, CA 94304-1050
Telephone: (650) 493-9300

Counsel for Defendant
GOOGLE LLC