[Counsel on signature page]

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL.,

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

Case Number: 5:24-cv-03811-EKL-SVK

JOINT DISCOVERY SUBMISSION REGARDING DISPUTE AS TO SAMPLING PROTOCOL TO ADDRESS PUBLISHERS' RFP NOS. 50-51

REDACTED

Judge Eumi K. Lee
Magistrate Judge Susan van Keulen

Pursuant to the Court's March 25, 2025 Order, Dkt. 318, Section 8 of the Court's Civil Standing Order, and L.R. 37-1 and 37-2, Plaintiffs ("Publishers") and Defendant Anthropic PBC ("Anthropic") respectfully submit this Joint Discovery Submission regarding the development of a sampling protocol to resolve the dispute regarding Publishers' RFPs 50-51. The Parties' lead counsel met and conferred via videoconference (including March 31 and April 14, 2025) and email, but the Parties are unable to reach agreement. The current fact discovery deadline is in 134 days.

I. <u>**Publishers' Position**</u>

Despite Publishers' proposing several different compromises, the Parties have unfortunately been unable to reach agreement on the size and components of a statistically significant sample of Claude prompts and output sufficient to address Publishers' RFPs 50-51.

Although Anthropic proposed to the Court at the Mar. 18 discovery hearing that the Parties "pick a big enough sample" so that "both sides can run whatever searches they want," Hr'g Tr. 17:20-21, 21:21-23 (Mar. 18, 2025), it has since tried to walk back that commitment considerably, seeking to shrink the sample's size and undermine its utility to Publishers. Anthropic's current proposal is not aimed at allowing Publishers a meaningful opportunity to search and analyze Claude prompts and output (including, in particular, the various lyric-related records requested by RFPs 50-51) or draw statistically significant conclusions from that analysis. Rather, Anthropic's proposal is aimed only at buttressing its own fair use defense. That is not the right approach.

Publishers' proposal, by contrast, is in line with Anthropic's prior commitment to produce a sufficiently large sample such that each party can analyze the data for multiple purposes. Specifically, Publishers propose that Anthropic produce a sample dataset of <u>24.5 million</u> Claude prompt and output records, or the rough equivalent of (1) [redacted] days between Sept. 22 and Oct. 18, 2023 (pre-lawsuit), and (2) [redacted] days between Oct. 19, 2023 and Mar. 22, 2024 (post-lawsuit). Such a dataset represents a modest [redacted]<u>% of the total population of</u> [redacted] <u>million records</u> Anthropic has preserved for this six-month period. As detailed below, Publishers' approach will ensure the sample is sufficiently large to draw statistically significant conclusions across a range of complicated issues important to both sides—not just Publishers' claims or Anthropic's defenses.

Anthropic's much more limited proposed sample—1 million records, a mere [redacted]% of the total records—is too small for the Parties to draw sufficiently reliable conclusions across the data.

A. Publishers' proposed sample size will allow statistically significant analysis and conclusions as to a range of different types of prompts and output.

At the outset, it bears emphasizing that, not only is the Claude dataset very large, but the data is complicated and the analyses varied, necessitating the larger sample size Publishers propose. Publishers' RFPs 50-51 seek Claude prompts and output that generally "relate to song lyrics," ECF No. 318 at 1, but within this category are myriad, overlapping sub-categories of relevant prompts and output, the precise details of which will be unknown until the Parties review the sample. Likewise, the Parties may analyze various other types of prompts and output as part of the sample.

Some types of Claude prompts and output to be analyzed may occur in relatively large numbers. For example, out of the [redacted] million total records at issue, upwards of 6.5 million contain the term "lyric." (While Anthropic has repeatedly refused to search for the term "lyric" across the full dataset, this estimate is based on Anthropic's representation that, for a nine-day period in Sept. 2023, the term "lyric" appeared in 170,077 of [redacted] million records—[redacted]***% of all records***.) That will include a wide range of prompts and output—such as prompts specifically requesting Publishers' and others' song lyrics, output copying lyrics even when the user did not ask, and records in which the term "lyric" refers to something other than song lyrics. There are also likely millions of additional records in which users seek or receive lyrics to Publishers' works and other songs, but the specific term "lyric" does not appear (*e.g.*, "What are the words to I Will Survive?", "What is the first verse to What a Wonderful World?", "Write me a song about the death of Buddy Holly").

At the other end of the spectrum, there are Claude prompts and output that—while much less frequent relative to the full dataset—are nevertheless vitally important to the case, such as:

- Instances in which one of ***Anthropic's own founders uses the AI model to seek lyrics***, *e.g.*, Anthropic_0000016458 (Anthropic co-founder Tom Brown querying, "@Claude what are the lyrics to desolation row by [Bob] Dylan?"); Am. Compl. ¶ 113 (alleging same); and
- Instances in which Claude copies Publishers' lyrics in output and then, when specifically asked by the user whether those lyrics are copyrighted, Claude ***falsely claims*** the lyrics are "entirely fictional and written by myself" and "free of any copyrighted material," *id.* ¶ 104.

Prompts and output like these—even if one in a million—are critically relevant to Publishers' claims and Anthropic's willful infringement. They must be accounted for in deciding a sample size.

This variance in the prevalence of prompts and output to be analyzed, the overlapping and subjective nature of the data, and the fact that much about the data to be analyzed remains unknown at this time all favor a larger sample size to ensure statistical significance. Ex. 1, Buchan Decl. ¶¶ 13-14, 19, 37. Publishers' proposed 24.5-million-record sample will better allow the Parties to analyze both frequent and infrequent occurrences alike, ensuring any conclusions derived from the sample are more reliable and precise. *Id.* ¶¶ 30-37; *In re Countrywide Fin. Corp. Mortg.-Backed Sec. Litig.*, 984 F. Supp. 2d 1021, 1033 (C.D. Cal. 2013) ("[T]o draw reliable conclusions about a population based on a statistical sample, the sample size must be large enough to support those conclusions."). The larger the sample, the more precise the conclusions. Buchan Decl. ¶¶ 14, 37.

By comparison, Anthropic's proposed 1-million-record sample is too small to account for the more rare occurrences—such as Anthropic's founder using Claude to search for Publishers' lyrics, the ***very thing Anthropic denied*** designing Claude to do—that are still central to the case.

Anthropic also seeks to apply an "assumed prevalence rate" of 0.006% across the board. But that one-size-fits-all approach is inappropriate here. As explained above, assuming any single prevalence rate vastly oversimplifies the many unique types of prompts and output and different issues to be analyzed. Publishers' larger sample size takes these considerations into account.

Publishers' proposed sample size is also more consistent with commonly accepted surveying and sampling practices. Publishers' sample size calculation is based on the following:

- <u>95% confidence level</u>: Publishers understand that the Parties are in agreement on this figure.
- <u>5% relative margin of error</u>: This margin of error is commonly utilized for sampling purposes, including rare event sampling, and comes with the added benefit of increasing the precision of any estimates derived from the sample data. Buchan Decl. ¶ 29.
- <u>Anthropic's "assumed prevalence rate" of 0.006%</u>: Publishers are willing to accept, *arguendo*, Anthropic's assumed prevalence rate for purposes of calculating a sample size.

Based on these figures, a statistically significant sample is 24.5 million records. *Id* ¶¶ 34, 36.

By contrast, Anthropic's proposed 1-million-record sample utilizes the same 95% confidence level and 0.006% prevalence rate, but is premised on a much larger and less precise 25% relative margin of error. A 25% margin of error is not appropriate here and is on the very outer threshold of standard practice. *Id.* ¶¶ 30-32. This inflated margin of error appears designed to reduce the sample's size at the cost of undermining the sample's utility. *See id.* ¶ 37.

B. Publishers' proposed approach will also allow statistically significant analyses of Claude prompts and output both pre-suit and post-suit.

It is also imperative that the sample allow the Parties to draw reliable conclusions about Claude prompts and output from both (1) before Publishers filed suit, when Anthropic's limited guardrails that were largely ineffective in preventing infringing output, and (2) after Publishers filed suit, when Anthropic adopted additional post-litigation guardrails in response to Publishers' claims. Accordingly, Publishers propose that half of the sample be drawn from Anthropic's [redacted] million Claude records from Sept. 22 to Oct. 18, 2023 (pre-lawsuit), and half from its [redacted] million records from Oct. 19, 2023 to Mar. 22, 2024 (post-lawsuit). This approach is consistent with the Court's guidance that "the sample must include both pre-suit and post-suit prompts and outputs." ECF No. 318 at 2. By contrast, Anthropic's position that prompts be drawn randomly from across the six-month period for which it preserved and agreed to produce records will skew the sample to the post-suit period, when Claude prompts and output may be very different than during the pre-suit period, in a number of ways. That Anthropic's retention policy did not provide for preserving records for a longer period pre-suit should not allow it to distort the sample in this manner.

C. Publishers' proposed sampling approach will not unduly burden Anthropic.

In addition to ensuring statistical significance, Publishers' proposal also reasonably balances Publishers' need for this discovery against the burden to Anthropic. While Anthropic may face a slightly greater burden in collecting and producing a larger dataset, it has "not provided any information that indicates that the production of such [records] would be unduly burdensome." *Fed. Trade Comm'n v. Tate's Auto Ctr. of Winslow Inc.*, 2019 WL 1130006, at *4 (D. Ariz. Mar. 12, 2019) (directing parties to identify an appropriate statistical sample size). That is particularly true where Anthropic will simply be required to process and produce the data at issue—which is minimally burdensome—rather than undertake any substantive review prior to production. Moreover, Anthropic has acknowledged its systems are capable of searching and processing at least [redacted] million Claude records at a time (the equivalent of nine days). Hr'g Tr. 19:7–13 (Mar. 18, 2025). Publishers' proposal would require only that Anthropic process 24.5 million records (the rough equivalent of [redacted] days, which can be broken up into smaller portions of nine days or less).

D. Anthropic's proposed approach will serve only its own ends.

Finally, it bears emphasizing that the reason the Parties are undertaking this sample in the first place is to address Publishers' RFPs 50-51, which seek production of Claude prompts and output relating to song lyrics. ECF No. 318 at 1. Over the course of the Parties' discussions regarding this sample, however, it has become increasingly apparent that Anthropic is seeking to exploit this sample as an opportunity to buttress its own fair use defense—by focusing on "overall Claude usage" beyond that at issue in Publishers' RFPs 50-51—while at the same time limiting production of the very lyrics-related records sought by Publishers' RFPs. This is improper. Anthropic cannot substitute its own objectives for Publishers' in this manner. *See, e.g.*, *Friedman v. 24 Hour Fitness USA, Inc.*, No. CV0606282AHMCTX, 2009 WL 10672797, at *1 (C.D. Cal. Jan. 12, 2009) ("It would be unfair to limit Plaintiffs to the 200–member sample provided by the Discovery Order and allow Defendant to use evidence regarding members outside the sample.").

Given that Publishers' RFPs 50-51 are the basis for this discovery dispute and the Court's order that the Parties develop a sampling protocol, and given that Publishers' RFPs 50-51 specifically seek Claude prompts and output relating to song lyrics, the sample that the Parties undertake must at a minimum allow for Publishers to identify and analyze a sufficient number of the specific lyric-related prompts and output they request. Each such prompt and output is a potentially new and separate act of infringement by Anthropic. Each is relevant to Publishers' claims of direct infringement, secondary copyright infringement, and removal of copyright management information, as well as Anthropic's willfulness, damages, and other core issues.

For this purpose, Publishers seek a reasonably-sized sample—24.5 million records—that balances Publishers' entitlement to the documents they seek against the burden of that discovery.

Critically, Publishers' proposed sample is sufficiently large to address ***both*** Publishers' discovery and Anthropic's goal of analyzing the sample for its own purposes, erring on the side of caution and with the objective of avoiding further discovery disputes as to this sample. Anthropic's too-small sample, on the other hand, is geared solely toward propping up its fair use defense.

In sum, only Publishers' proposal will fairly allow both Parties to "run whatever searches they want," as Anthropic initially committed, and "harvest the necessary information from the dataset in an efficient, effective, and timely manner," as the Court ordered, ECF No. 318 at 2.

II. Anthropic's Position

Sampling reduces the burden and expense of discovery regarding voluminous data. *See* Manual for Complex Litigation § 11.493 (4th ed. 2004) ("Acceptable sampling techniques, in lieu of discovery and presentation of voluminous data from the entire population, can save substantial time and expense, and in some cases provide the only practicable means to collect and present relevant data."); *accord Tyson Foods, Inc. v. Bouaphakeo*, 577 U.S. 442, 454–55 (2016) (approving use of "a representative or statistical sample" as a "means to establish or defend against liability"). Plaintiffs sought broad discovery of all Claude prompts and outputs related to song lyrics within a dataset of approximately [redacted] records. This Court recognized the "technical infeasibility and burden" of that request and adopted Anthropic's position that a random sample would "allow the Parties to harvest the necessary information from the dataset in an efficient, effective and timely manner." Dkt. 318 at 2; *see id.* (ordering Anthropic to "produce a statistically significant sample of the Claude prompt and output records" after meeting and conferring about the sample's "the size and components"). Anthropic's proposed sample of 1 million records is more than sufficient to evaluate actual Claude usage patterns. Plaintiffs' proposal—to simply select [redacted] full days of data (over 10 million records)—defies basic statistical principles and flatly contradicts their theory of the case. If users were truly exploiting Claude to reproduce lyrics as often as Plaintiffs claim, Plaintiffs would need only a fraction of this data to prove it.

When the parties met and conferred on March 31, 2025, Plaintiffs proposed the use of a "pre-sample sample" to develop a rough estimated prevalence rate for lyric-related prompts and outputs. Under this approach, the parties would have used the estimated prevalence rate from a pilot study of approximately 200,000-260,000 records to calculate, using standard statistical formulas, an appropriate sample size for the entire population. Statistical principles dictate that a small estimated prevalence rate from the pilot study merits a larger sample size, and vice versa. The parties negotiated the details of this pre-sample sample in good faith between March 31 and

April 9, at which point Plaintiffs abruptly abandoned this approach and instead proposed, for the first time and without any mathematical rationale, a sample of [redacted] full days of data. This proposal would have resulted in a sample size of over 20 million records. Since then, Plaintiffs have modified their proposal to [redacted] full days of data—[redacted] days pre-suit and [redacted] days post—for a sample size still totaling over 10 million records. This proposal is neither scientific nor reasonable.

Anthropic has proposed a random sample of 1 million records. Under Anthropic's proposal, 1 million chat logs (i.e., user interactions containing both prompts and outputs) would be selected randomly from across the 6-month universe of data. This sample is large enough to allow meaningful conclusions about the entire dataset, but small enough to significantly reduce the burden and expense of discovery into song-related prompts and outputs. Plaintiffs' proposal of a fixed sample of 10+ million records selected from [redacted] days of data is flawed in two respects: (1) it is far larger than necessary to derive meaningful insights about the population in question, either because it implicitly assumes an incomprehensibly low prevalence rate or uses a needlessly precise margin of error; and (2) its focus on just [redacted] days of data—[redacted] from before the complaint was filed and [redacted] after—injects selection bias and is thus unlikely to be representative.

A. Anthropic's Sample Size of 1 Million Is An Appropriate Size to Draw Conclusions Regarding the Overall Population

In order to realize the efficiencies of statistical sampling, a sample should be no larger than necessary to achieve statistical validity. *See* Declaration of Qinnan (Olivia) Chen ("Chen Decl."), Ex. B ¶ 18. Sample size is typically calculated based on an estimated prevalence in the population of the event being studied and the relative margin of error. *Id.* ¶ 5. If an expected prevalence rate is unavailable, either a pilot study can be conducted to determine one, or a "guesstimate" can be used.[1] Here, the parties' earlier search term negotiations and prompt and output productions thus far demonstrate the relative rarity of the event being studied (prompts for or outputs containing song lyrics), *see* Dkt. 302 at 6–7, 9, but not the precise prevalence rate.

In the absence of a pilot study, Anthropic's 1 million record sample is sufficiently representative to reflect even extremely rare events. *See* Chen Decl. ¶¶ 6–8. Using a 25% relative

[1] *See* Chittaranjan Andrade, *Sample Size and Its Importance in Research*, 42 Indian J. Psych. Med. 102, 103 (2020), *available at* https://pmc.ncbi.nlm.nih.gov/articles/PMC6970301/.

margin of error—widely accepted for extremely rare events—a sample size of 1 million would validly reflect events in the overall population with prevalence rates as low as 0.006%. Using a more stringent 20% relative margin of error, a 1 million record sample size would still capture events with prevalence rates as low as 0.01%. *Id.* ¶¶ 9–12. Any statistical power gained from a larger sample would be outweighed by the diminished efficiencies involved in running and reviewing the results from a larger sample. *Id.* ¶ 17–18; *see also* Fed. R. Civ. P. 26(b)(1) (courts must consider if "the burden or expense of the proposed discovery outweighs its likely benefit").

To the extent Plaintiffs argue that a sample of 1 million is inadequate because it represents too small a fraction of the overall population, that argument is unmoored from rigorous and accepted statistical methods. *See* Chen Decl. ¶ 5 (setting forth the formula that "represents the fundamental statistical approach to determining the minimum sample size needed to make valid inferences about a very large dataset," including "the one at issue here"). It is also unsupported by case law in this Circuit, which recognizes that the fraction of a representative sample should decrease as the total population (and burden) increases. *See Guzman v. Chipotle Mexican Grill, Inc.*, 2018 WL 6092730, at *3 (N.D. Cal. Nov. 21, 2018) (finding a 5% sample appropriate for a "large" population of 43,000 and noting that courts imposed higher percentage samples for "much smaller" populations); *Heredia v. Sunrise Senior Living LLC*, 2019 WL 7865176, at *6 (C.D. Cal. Oct. 31, 2019) (finding a 15% sample of a 13,000 population "appropriately balances Plaintiffs' discovery needs, Defendant's burdens, and the privacy rights" of the individuals sampled and noting that a larger sample "would be excessive"); *Nia v. Bank of Am., N.A.*, 2023 WL 2583386, at * 5 (S.D. Cal. Mar. 20, 2023) (considering the significant costs required to "process, review, redact, and produce the proposed sample" in deciding to cut the sample size from 20% to 10%).

Plaintiffs' insistence on a sample *ten times* the size that Anthropic proposes cannot be supported by accepted statistical methods. *See Deutsche Bank Nat'l Tr. Co. v. Morgan Stanley Mortg. Cap. Holdings LLC*, 289 F. Supp. 3d 484, 496 (S.D.N.Y. 2018) ("Properly done, statistical sampling is not guesswork"). A sample of this size either assumes (a) an extraordinarily low prevalence rate, far lower than the 0.01% or 0.006% already contemplated by Anthropic's sample; or (b) an unnecessarily stringent relative margin of error. *Id.* ¶ 17. With respect to (a), Plaintiffs' position throughout this litigation has been the opposite: that Claude was designed to—and frequently does—respond to prompts requesting song lyrics. *See, e.g.*, Dkt. 1 ¶ 11 ("Anthropic unlawfully enables, encourages, and profits from massive copyright infringement by its users"); Dkt. 179 at 2 ("Anthropic intended and expected its AI models to respond to requests for Publishers' lyrics—as a feature, not a bug"); *and* Dkt. 225 at 10 (pointing to search results for the terms "lyric" and "song" to posit a "potentially huge volume of requests for lyrics by [Claude] users"). In their amended complaint filed last week, Plaintiffs allege that "literally millions of Claude prompt and output records contain the term 'lyric,'" and that this means that "*[c]ountless* users . . . have prompted Claude for lyrics." Dkt. 337 ¶¶ 9–10 (emphasis added). Plaintiffs now conveniently treat lyric requests as extraordinarily rare events comparable to rare genetic diseases or lightning strikes. If their claims have merit (which Anthropic disputes), Anthropic's sampling proposal is already more than adequate. And with respect to (b), Plaintiffs cannot explain why the Court must exceed the margin of error standards that are routinely accepted by statisticians when analyzing rare events. Chen Decl. ¶¶ 9, 17. Plaintiffs' proposed sample size should be rejected.

B. Anthropic's Proposed Sample Is Representative of the Entire Population

To extrapolate the results of a sample to the larger population, the sample "must be randomly selected." *United States v. Nunez*, 2021 WL 5494588, at *7 (D. Conn. Nov. 23, 2021). "Random" means that every datapoint in the population has the same probability of becoming part of the sample. *See* D. Freedman & D. Kaye, FJC, *Manual on Scientific Evidence* 230 (2011). Scientific randomness "provides assurance of unbiased estimates." *Id*; *see also* Chen Decl. ¶ 19.

Anthropic has proposed a true random sample: each Claude user interaction in the full dataset would have an equal chance of being included. This would ensure the sample includes both pre- and post-suit chat logs but does not suffer from any biases resulting from hand selecting chat logs from non-representative periods, e.g., (1) weekends versus weekdays, or vice versa; (2) months when Claude had a smaller userbase; and (3) periods of anomalous user activity (i.e., around the filing of Plaintiffs' complaint when their agents were "investigating"). Chen Decl. ¶¶ 19–22. Anthropic's proposed sample would be unbiased and representative of the full population.

Plaintiffs' proposal is not random in the rigorous, scientific sense because it suffers from each of the above temporal biases. *Id.* "Hand-picking is almost certain to introduce a substantial amount of selection bias into the sample," *Rosenbohm v. Cellco P'ship*, 2019 WL 2141901, at *2 (S.D. Ohio May 16, 2019), *objections overruled*, 2019 WL 13507817 (S.D. Ohio July 24, 2019), and Plaintiffs can offer no rationale, grounded in established statistical methods, for introducing such biases, *see also* Freedman & Kaye, 230 ("Looser definitions of randomness are inadequate for statistical purposes."). Their proposal is unlikely to produce representative results that would allow extrapolation to the larger population and should be rejected.

C. Regardless of the Contours of the Sample, Privacy Safeguards Are Needed

The parties also disagree on the privacy protections required in connection with this sample. These chat logs represent the private communications of third-party Anthropic users. *See generally* Dkt. 310 at 7–11. Unlike the prompts and outputs that Anthropic has produced thus far, however, the prompts and outputs in the sample will not be filtered for relevance. Accordingly, the privacy interests of the users associated with these chat logs are exceptionally strong. *Cf. Heredia*, 2019 WL 7865176, at *4 (expressing "concern[] about the privacy interests" of the individuals swept into a sample, and balancing those interests in ordering a sampling procedure).

To adequately protect these interests, Anthropic proposed—and Plaintiffs materially agreed to—a set of modest procedures, including: (1) that the sample prompts and outputs as well as any aggregate information and/or analysis would be designated Confidential or HC-AEO under the Protective Order, subject to Plaintiffs' right to challenge those designations; (2) that the sample prompts and outputs would be de-identified, but to the extent any identifying information was not removed from the record prior to disclosure, Anthropic would be given a reasonable opportunity to remove it; and (3) that the sample would be accessible only on a secure platform.

Despite Plaintiffs' initial assent to a set of privacy-protecting procedures, they withdrew that assent when Anthropic would not agree to their' proposal of [redacted] full days of data. Anthropic respectfully requests that the Court incorporate these protections into a sampling protocol.

Dated: April 30, 2025

By: *<u>/s/ Timothy Chung</u>*
OPPENHEIM + ZEBRAK, LLP
Matthew J. Oppenheim
Nicholas C. Hailey
Audrey L. Adu-Appiah
(admitted *pro hac vice*)
4530 Wisconsin Ave., NW, 5th Floor
Washington, DC 20016
Telephone: (202) 480-2999
matt@oandzlaw.com
nick@oandzlaw.com
aadu-appiah@oandzlaw.com

Jennifer Pariser
Andrew Guerra
Timothy Chung
(admitted *pro hac vice*)
461 5th Avenue, 19th Floor
New York, NY 10017
Telephone: (212) 951-1156
jpariser@oandzlaw.com
andrew@oandzlaw.com
tchung@oandzlaw.com

COBLENTZ PATCH DUFFY & BASS LLP
Jeffrey G. Knowles (SBN 129754)
One Montgomery Street, Suite 3000
San Francisco, CA 94104
Telephone: (415) 391-4800
ef-jgk@cpdb.com

COWAN, LIEBOWITZ & LATMAN, P.C.
Richard S. Mandel
Jonathan Z. King
Richard Dannay
(admitted *pro hac vice*)
114 West 47th Street
New York, NY 10036-1525
Telephone: (212) 790-9200
rsm@cll.com
jzk@cll.com
rxd@cll.com

Attorneys for Plaintiffs

Respectfully submitted,

By: *<u>/s/ Sarang V. Damle</u>*
LATHAM & WATKINS LLP
Joseph R. Wetzel (SBN 238008)
joe.wetzel@lw.com
Andrew M. Gass (SBN 259694)
andrew.gass@lw.com
Brittany N. Lovejoy (SBN 286813)
britt.lovejoy@lw.com
505 Montgomery Street, Suite 2000
San Francisco, California 94111
Telephone: +1.415.391.0600

Sarang V. Damle
(admitted *pro hac vice*)
sy.damle@lw.com
Sara Sampoli (SBN 344505)
sara.sampoli@lw.com
555 Eleventh Street NW, Suite 1000
Washington, DC 20004
Telephone: +1.202.637.2200

Allison L. Stillman
(admitted *pro hac vice*)
alli.stillman@lw.com
1271 Avenue of the Americas
New York, New York 10020
Telephone: +1.212.906.1747

Attorneys for Defendant

SIGNATURE ATTESTATION

Pursuant to Civil L.R. 5-1(i)(3), I hereby attest that concurrence in the filing of this document was obtained from all other signatories of this document. I declare under penalty of perjury that the foregoing is true and correct.

Dated: April 30, 2025

/s/ Timothy Chung
Timothy Chung

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

NOTICE

This document has been removed by order of the court.

For more information, please see the entire docket sheet, or contact the clerk's office, or consult chambers.

LATHAM & WATKINS LLP
Joseph R. Wetzel (SBN 238008)
joe.wetzel@lw.com
Andrew M. Gass (SBN 259694)
andrew.gass@lw.com
Brittany N. Lovejoy (SBN 286813)
brittany.lovejoy@lw.com
Ivana Dukanovic (SBN 312937)
ivana.dukanovic@lw.com
505 Montgomery Street, Suite 2000
San Francisco, California 94111
Telephone: +1.415.391.0600

Sarang V. Damle (*pro hac vice*)
sy.damle@lw.com
Sara Sampoli (SBN 344505)
sara.sampoli@lw.com
555 Eleventh Street NW, Suite 1000
Washington, D.C. 20004
Telephone: +1.202.637.2200

Allison L. Stillman (*pro hac vice*)
alli.stillman@lw.com
1271 Avenue of the Americas
New York, New York 10020
Telephone: +1.212.906.1747

Rachel Horn (SBN 335737)
rachel.horn@lw.com
140 Scott Drive
Menlo Park, California 94025
Telephone: +1.650.328.4600

Attorneys for Defendant Anthropic PBC

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL.,

Plaintiffs,

vs.

ANTHROPIC PBC,

Defendant.

Case No. 5:24-cv-03811-EKL-SVK

DECLARATION OF OLIVIA CHEN IN SUPPORT OF ANTHROPIC'S SAMPLING PROPOSAL IN CONNECTION WITH JOINT DISCOVERY DISPUTE

Hon. Eumi K. Lee
Magistrate Judge Susan van Keulen

1. My name is Qinnan (Olivia) Chen, and I am a Data Scientist at Anthropic, PBC. I submit this declaration in support of Anthropic's sampling proposal in connection with the pending Joint Discovery Dispute Statement. Dkt. 318. Unless stated otherwise, all facts stated herein are within my personal knowledge. If called upon, I would and could competently testify as to matters contained in this declaration.

2. I understand that on March 25, 2025, the Court ordered Anthropic to produce a "statistically significant" sample of Claude.ai prompt and output records from a dataset of hundreds of millions of records spanning from September 22, 2023 to March 22, 2024.[1] I further understand that, at a minimum, the Court stated that the sample must include both pre-suit and post-suit prompts and outputs and must not separate the outputs from their prompts. I understand that despite extensive efforts to reach an agreement on a sampling protocol, the parties have been unable to find common ground and are therefore submitting their respective positions regarding the appropriate sample size and methodology for establishing a statistically significant sample.

3. I hold a Bachelor's Degree in Economics and Communication from the University of California, Davis and a Master's Degree in Statistics from American University. I have worked as a data scientist for almost nine years, and have received certifications in the following: dbt Fundamentals, Neural network and Deep Learning, and SAS Certified Base Programmer for SAS 9.

4. Because of my educational and professional background, I am very familiar with the well-established methodologies for drawing representative samples from which reliable conclusions about a larger population can be drawn. When determining an appropriate sample size, statisticians rely on several key techniques, including: simple random sampling, stratified sampling, cluster sampling, and systematic sampling.

[1] In the field of statistics, the term "statistical significance" typically relates to the result of a hypothesis test—e.g., evaluating whether an observed effect in data is likely due to something other than random chance. The term is not typically used to describe a sample of data itself. But I understand the Court to have essentially ordered the production of a "representative" sample—*i.e.*, sample of sufficient size to accurately estimate the prevalence of the relevant event (users seeking lyrics) in the full dataset.

5. The foundation of these approaches is the sample size formula, which is calculated based on several factors including the expected prevalence of the phenomenon being studied. For very large datasets, the formula is:

$$n = \frac{Z^2 \cdot (1 - p)}{E_{rel}^2 \cdot p}$$

- n = required sample size
- Z = Z-score (standard score) corresponding to the desired confidence levels (1.96 for 95% confidence)
- p = expected prevalence (or proportion of the event in the population)
- E_{rel} = relative margin of error, expressed as a proportion

This formula represents the fundamental statistical approach for determining the minimum sample size needed to make valid inferences about a very large dataset (like the one at issue here) with a specified level of confidence and precision.[2]

6. I understand the specific phenomenon under consideration involves an exceptionally rare event: the incidence of Claude users requesting song lyrics from Claude. I understand that this event's rarity has been substantiated by manual review of a subset of prompts and outputs in connection with the parties' search term negotiations and the prompts and outputs produced to date. In the absence of a pilot sample to calculate an estimated prevalence rate, a reasonable prevalence rate for a rare event could easily be as low as 0.01% of all user interactions.

I. Anthropic's Sampling Proposal for Prompt and Output Data

7. Based on established statistical principles and peer-reviewed research, Anthropic proposes a random sample of 1 million Claude.ai prompt and output records, equally distributed across the relevant time period from September 22, 2023, to March 22, 2024. This simple sampling technique will result in a comprehensive sample that will include both pre-litigation and post-

[2] *See, e.g.,* Penn State Univ., STAT 200: Elementary Statistics, *Sample Size Estimation*, https://online.stat.psu.edu/stat200/lesson/8/8.1/8.1.1/8.1.1.3 (last visited Apr. 30, 2025).

litigation interactions, as the lawsuit was initiated on October 18, 2023, and will maintain the integrity of the dataset by preserving prompt-output pairs as complete units.

8. Given the effectively unlimited nature of the dataset in question and the extremely low prevalence rates discussed above, statistical analysis confirms that a 1 million record sample size far exceeds what would be required to obtain a sample of sufficient size to draw accurate inferences about the prevalence of even rare events like seeking song lyrics. As demonstrated in my calculations below, this sample size provides exceptional confidence levels and minimal margins of error.

9. Using standard statistical methods, including the validated sample size formula outlined above, I have calculated that 614,595 prompt-output records would adequately capture a statistically significant cross-section of the relevant data for prevalence rates as low as 0.01% using a 25% relative margin of error. This 25% relative margin of error is widely accepted by statisticians as reasonable and appropriate when estimating sample sizes for extremely rare events. Reliance on the 25% relative margin of error parameter is extensively supported by peer-reviewed research in medical statistics, epidemiology, and large-scale data analysis, where rare event detection must balance statistical power with practical limitations.[3]

10. Even if we apply more stringent statistical parameters than typically required for rare events like seeking song lyrics on Claude, an appropriate sample size would still be less than 1 million records. Based on calculations using the standard sample size formula, I have determined that 960,304 prompt and output records would be adequate to capture a statistically significant cross-section of the relevant data for prevalence rates as low as 0.01% using a more conservative

[3] *See* Julien Dutant & Julia Staffel, *A Statistician's Guide to Making Sound Inferences from Noisy Data*, 78 American Statistician 437, 437–449 (2024), https://www.tandfonline.com/doi/full/10.1080/00031305.2024.2350445; Lokesh K. Singh et al., *Brief Intervention for Tobacco when Diagnosed with Oral Cancer (BITDOC): Study protocol of a randomized clinical trial studying efficacy of brief tobacco cessation intervention, Chhattisgarh, India* at 4 (2020), https://pmc.ncbi.nlm.nih.gov/articles/PMC7291894/; Lower Windward Environmental LLC, *Lower Duwamish Waterway Pre-Design Studies Data Evaluation Report (Task 6)* at 6, 65 (2020), https://semspub.epa.gov/work/10/100248737.pdf.

20% relative margin of error. These calculations demonstrate that Anthropic's proposed sample size provides robust statistical power even under more demanding precision requirements.

11. I have further analyzed scenarios where the prevalence rate of song lyrics requests might be even lower than initially estimated. Notably, across multiple statistical scenarios with varying prevalence rates and confidence parameters, the mathematically sound sample size consistently converges around 1 million records.

12. For example, assuming an *extremely* low prevalence rate of 0.006% while maintaining the statistically accepted 25% relative margin of error would result in a required sample of 1,024,365 prompt and output interactions. This calculation, consistent with established statistical principles for rare event detection, further confirms that a sample of approximately 1 million records provides more than a statistically sound dataset from which to draw reliable conclusions about Claude usage patterns, including rare events such as lyrics requests.

13. A sample size of 1 million prompt and output interactions is also strategically sufficient to neutralize potentially confounding variables that must be accounted for to ensure statistical validity and representativeness. Anthropic's proposed 1 million record sample effectively controls for temporal variations in Claude interaction patterns—ensuring adequate representation of both high and low traffic periods across different days of the week and times of day. It would also successfully neutralize variations in user demographics, including subscriber status (paid versus free Claude users), geographic distribution, and language preferences, thereby providing a genuinely representative cross-section of the overall data population which amounts to hundreds of millions of records.

14. Anthropic's proposed 1 million record sample not only satisfies but substantially surpasses the requirement to produce a representative sample of Claude.ai interactions. It reflects statistical best practices for analyzing rare events within large-scale datasets and will provide a scientifically valid basis for drawing conclusions about the broader population of prompt-output interactions.

II. Publishers' Sampling Proposal for Prompt and Output Data

15. I understand that the Publishers have proposed various approaches during the parties' negotiations. Initially, I understand that the Publishers proposed a "pre-sample sample" methodology—or pilot sample—to determine the frequency with which Claude users request lyrics based on the population of data, which would then inform the calculation of an appropriate sample size using standard statistical methods. In other words, this "pre-sample sample" would have assisted in more precisely calculating the prevalence input for the sample size formula. At a minimum, this approach acknowledged the need for statistical rigor in determining sample parameters.

16. I understand that the Publishers subsequently abandoned this pre-sample sample approach and instead demanded the production of [redacted] complete days of prompt and output records ([redacted] days preceding and [redacted] days following the filing of the complaint). This revised proposal would have necessitated the production of over 20 million prompt and output records without any statistical justification or analysis. I further understand that the Publishers then revised their proposal again to request a sample of prompt and output interactions consisting of [redacted] full days of data (approximately 10 million records) from [redacted] days before and [redacted] days after the complaint was filed. I understand the Publishers have not provided the statistical basis for their newest proposal.

17. Both of these proposals represent extreme outliers in statistical practice for sampling rare events and are unnecessary to analyze typical Claude usage. Such large samples would be unnecessary except where the prevalence rate is incomprehensibly low, which I understand is contrary to positions the Publishers have taken elsewhere in this litigation. One alternative explanation for such a large sample size would be the use of an unnecessarily stringent relative margin of error. There is an inverse relationship between prevalence and relative margin of error, which means that a more stringent relative margin of error for a rare event requires an enormous sample size. But there are diminishing benefits to such large samples, since the marginal improvement in the absolute margin of error would be incredibly small. A sample size of either 10

or 20 million is not necessary or advisable to achieve statistically valid results for even very rare events.

18. This is because a sample that is larger than necessary risks diminishing returns; any potential benefit would be significantly outweighed by the effort and expense required to properly analyze such a large dataset, especially where a 1 million record sample would be considered sufficient. A larger sample also requires and consumes more resources. In the field of statistics, it is considered an unethical waste of resources to use unnecessarily large samples.

19. Both variations of the Publishers' sampling proposal also suffer from fundamental methodological flaws that would severely compromise the statistical validity of any findings derived from such samples. First, data collected exclusively from a fixed set of calendar days before and after the complaint presents significant risks of temporal bias and would fail to be representative of the entire universe of interactions across the relevant time period (September 22, 2023 to March 22, 2024). This systematic bias would produce distorted results that could not be reliably extrapolated to the broader population of interactions. In contrast, proper random sampling techniques across the entire time period, as proposed in Anthropic's methodology, would effectively eliminate this source of bias while requiring only a fraction of the data volume.

20. Second, the Publishers' proposed fixed-day sampling method lacks the diversity of a wider time window, and introduces multiple additional sources of non-representativeness that would further undermine statistical validity. These include, for instance: (1) day-of-week biases that fail to account for documented variations in user behavior between weekdays and weekends; (2) failure to account for Anthropic's rapidly evolving user base during the relevant period; (3) heightened risk of capturing anomalous activity in the days immediately surrounding the legal filing, including potential testing or monitoring by Publishers or their agents that would not represent typical user behavior; and (4) failure to account for product updates or marketing campaigns that may have influenced user behavior during the selected timeframe.

21. In sum, fixed-day sampling is a high-volume, high-cost method that risks introducing biases that would not be present in a diverse sample from a wider time window. A

smaller, true random sample can achieve superior statistical results in a more cost-effective and efficient way.

22. Based on my professional expertise, I find that the Publishers' sampling proposal lacks scientific validity, contradicts established statistical principles for representative sampling, and would impose an unnecessary burden without corresponding analytical benefits.

23. Anthropic's proposed sample size of 1 million records strikes the reasonable balance between statistical power and analytical practicality. A smaller sample than that proposed by Anthropic would be statistically valid for the reasons above. It is a conservative approach to account for the possibility that the events in question are even rarer. In contrast, an unnecessarily larger sample such as that proposed by Publishers would introduce significant inefficiencies without corresponding statistical benefits. Excessive sample sizes can overwhelm analytical resources, dramatically increase processing time, and introduce needless computational complexity—all without materially improving statistical confidence or precision. Statistical principles dictate that once a sample size reaches the threshold of representativeness, additional sampling yields rapidly diminishing returns. Anthropic's proposed 1 million record sample achieves this equilibrium point, providing robust statistical validity while remaining practically manageable for thorough expert analysis.

I declare under penalty of perjury that to the best of my knowledge, information, and belief, the foregoing statements are true and correct.

Executed on April 30, 2025 in San Francisco, California.

Dated: April 30, 2025

Olivia Chen