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.

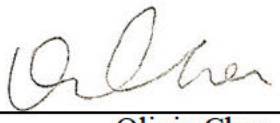
22. Based on my professional expertise, I find that the Publishers' sampling proposal lacks scientific validity, contradicts established statistical principles for representative sampling, and would impose an unnecessary burden without corresponding analytical benefits.

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