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) 772-5795
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

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

CONCORD MUSIC GROUP, INC., ET AL., Plaintiffs, v. ANTHROPIC PBC, Defendant.	Case Number: 3:24-cv-03811-JSC **DECLARATION OF ED NEWTON-REX IN SUPPORT OF PLAINTIFFS' MOTION FOR PRELIMINARY INJUNCTION** Hon. Jacqueline Scott Corley

I, Ed Newton-Rex, hereby declare pursuant to 28 U.S.C. § 1746:

1. I am the CEO and founder of Fairly Trained, a non-profit organization that certifies generative AI ("GAI") companies that obtain licenses for the data they use or use public domain or openly licensed data.

2. I submit this declaration in connection with Plaintiffs' Motion for Preliminary Injunction filed by Plaintiffs Concord Music Group, Inc., Capitol CMG, Inc., Universal Music Corp., Songs of Universal, Inc., Universal Music – MGB NA LLC, Polygram Publishing, Inc., Universal Music – Z Tunes LLC, and ABKCO Music, Inc. (collectively, "Publishers"), in their lawsuit against Defendant Anthropic PBC ("Anthropic").

3. My statements set forth below are based on my specialized knowledge, education, and experience, as applied to the facts and circumstances of this case. If called upon, I would and could competently testify as to the matters contained herein.

I. My Assignment

4. The Publishers have asked me to consider the factors influencing whether or not a GAI model developer would choose to seek authorization from copyright holders for data it uses to train its AI models, and the market for data licenses or public domain data in the GAI business.

5. In undertaking this work, I reviewed a variety of materials including relevant parts of both Parties' submissions in connection with Publishers' Motion for Preliminary Injunction and supporting declarations concurrently filed with this declaration, as well as press reports describing developments in the GAI industry and recent agreements between GAI developers and content companies.

II. My Background

6. I am the CEO of Fairly Trained, which I founded in 2024. Fairly Trained is a non-profit organization that certifies organizations that meet our requirements for data provenance for GAI models. We issue the Licensed Model (L) Certification to any company, product, or model for which its training data is licensed to or owned by the developer or is publicly available under an open license or is in the public domain; and where the organization has a robust due diligence process to confirm the source of its data and keeps records of their data sources. More information can be found on our website.[1]

7. In 2012, I founded Jukedeck, one of the first GAI companies, which provided music for video, TV, radio, podcasts and games. I created Jukedeck because I believed that AI could have

[1] FAIRLY TRAINED, https://www.fairlytrained.org/ (last visited July 31, 2024).

many positive use cases in music creation, including as a powerful tool to assist media creators with the creation of background music and helping aspiring songwriters in creating original music. Users could set parameters for the composition including genre, instrument selection, length and other features. Jukedeck would then generate a song in around a minute or less, which the composer could further edit to suit their purposes. Jukedeck was used to create over 1 million pieces of music, including for well-known brands such as Coca-Cola and Google, and won several awards, including an Innovation Lion at the 2016 Cannes Lions.

8. From 2019 to 2021, I was a Product Director at ByteDance, first in the AI Lab and then in the TikTok team. Most recently I was VP of Audio at Stability AI, where my team built and released Stable Audio, an AI music generation product that was named one of TIME's 100 Best Inventions of 2023. I resigned from Stability AI because I disagreed with the company's opinion that training generative AI models on copyrighted works is fair use. I explained my reasoning in a piece published in Music Business Worldwide, which is attached as **Exhibit B**.

9. I am a classically trained musician and compose choral music, which has been published by Novello and Boosey & Hawkes, a Concord Company. None of my compositions are listed as works in suit in this litigation. I was a judge in the first international AI Song Contest in 2020, and I am a mentor at Abbey Road Studios' music tech incubator.

10. I have also co-published several papers and patents concerning AI music generation.

11. A copy of my full CV is attached as **Exhibit A.**

III. Generative AI Model Training

A. Data Quality

12. As Dr. Ben Zhao explains in his declaration, generative AI models, including large language models ("LLMs"), are "trained" using machine learning algorithms, on datasets comprised of the type of content they are designed to output, such as text or audio files. After training, they can be used to generate content in connection with prompts by a given user.[2]

13. Unsurprisingly, the quality and efficacy of a GAI model is then greatly, but not wholly, affected by the quality of the data used to train the model. A training dataset comprised of

[2] Declaration of Ben Zhao ("Zhao Decl.") ¶ 13–15.

inaccurate information—for example a dataset in which references to the moon insist that it is made of Swiss cheese—would likely cause a GAI model trained on that data to output that same inaccurate information. Similarly, a model trained on copyrighted materials may output elements of those same copyrighted materials.[3]

14. The size of a model's training corpus also impacts performance, as Anthropic Chief Science Officer Jared Kaplan observed back in 2020 when he and his fellow authors concluded that larger language models trained on more data performed better and more efficiently.[4] In the time since, AI companies have largely determined to seek as much data as possible, utilizing datasets comprised of trillions of words to create their newest and most capable models.[5]

15. With publicly available data on the internet projected to dry up as soon as 2026,[6] AI companies are constantly on the hunt for new sources of high-quality content to train their latest models. For these reasons, libraries of copyrighted content, such as Publishers' lyrics, offer AI companies a valuable source of clean, organized data on which to build their models and enhance their capabilities.

B. Sources of Data

16. AI companies can source their data in a number of ways. The most common method is web scraping, by which AI companies like Anthropic will employ crawlers to scour the internet and obtain information considered "publicly available." This generally includes vast quantities of content that may be accessed from publicly available online sources but that is otherwise restricted from further use by relevant rightsholders. Thus, one of the fundamental issues with web scraping is that it all but guarantees that copyrighted materials will end up in the training corpus, unless the copyrighted materials are removed before training.

---

[3] *Id.* at ¶ 23–26.
[4] *See* Jared Kaplan et al., *Scaling Laws for Neural Language Models* (Jan. 23, 2020), arXiv:2001.08361v1, https://arxiv.org/pdf/2001.0836.
[5] Cade Metz et al., *How Tech Giants Cut Corners to Harvest Data for A.I.*, N.Y. TIMES (Apr. 6, 2024), https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.
[6] *See* Deepa Seetharaman, *For Data-Guzzling AI Companies, the Internet is Too Small*, WALL STREET JOURNAL (Apr. 1, 2021), https://www.wsj.com/tech/ai/ai-training-data-synthetic-openai-anthropic-9230f8d8.

17. Web scraping by AI companies also comes with negative ramifications on the accessibility of the internet, generally. In their efforts to draw from a dwindling source of publicly available data, several AI companies, including Anthropic,[7] have reportedly ignored website terms and conditions or robots.txt instructions prohibiting unauthorized data scraping on a given website.[8] Consequently, publishers and online platforms have increasingly restricted access to data on their websites, including by instituting paywalls, with the effect of making the internet more restrictive for all other users.[9] Large-scale crawling by Anthropic and other AI companies also places additional burdens on a website.[10]

18. Anthropic competitor OpenAI has reportedly transcribed millions of hours of YouTube videos for additional training data—conduct that is prohibited by YouTube's terms of service.[11] Google also reportedly tapped the same resource, as well as updating its terms of service to allow it to train AI models on publicly available content generated by its users on applications like Google Docs and Google Sheets.[12]

19. Like its competitors, there is evidence suggesting Anthropic also uses data collected in violation of YouTube's terms of service. Specifically, Anthropic has admitted to using data sourced from a dataset compiled by EleutherAI called the Pile,[13] which a recent investigation by

---

[7] Jess Weatherbad, *Anthropic's crawler is ignoring websites' anti-AI scraping policies*, VERGE (Jul. 25, 2024), https://www.theverge.com/2024/7/25/24205943/anthropic-ai-web-crawler-claudebot-ifixit-scraping-training-data.

[8] Dhruv Mehrotra and Tim Marchman, *Perplexity Is a Bullshit Machine*, WIRED (Jun. 19, 2024), https://www.wired.com/story/perplexity-is-a-bullshit-machine/.

[9] Kevin Roose, *The Data That Powers A.I. Is Disappearing Fast*, N.Y. TIMES (Jul. 19, 2024), https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html; Bobby Allyn, *Artificial intelligence web crawlers are running amok*, NPR (Jul. 5, 2024), https://www.npr.org/2024/07/05/nx-s1-5026932/artificial-intelligence-web-crawlers-are-running-amok.

[10] Jess Weatherbad, *Anthropic's crawler is ignoring websites' anti-AI scraping policies*, VERGE (Jul. 25, 2024), https://www.theverge.com/2024/7/25/24205943/anthropic-ai-web-crawler-claudebot-ifixit-scraping-training-data.

[11] Cade Metz et al., *How Tech Giants Cut Corners to Harvest Data for A.I.*, N.Y. TIMES (Apr. 6, 2024), https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html

[12] *Id.*

[13] Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment* 8, 27, ANTHROPIC (Dec. 9, 2021), https://arxiv.org/pdf/2112.00861.pdf.

Proof News revealed contained subtitles harvested from 173,536 YouTube videos across 48,000 channels—in violation of the platform's rules[14]—including subtitles to official music videos containing Publishers' lyrics.[15]

20. AI companies can procure datasets by other means, including by purchasing or obtaining them from dataset compilers. This includes widely used and freely accessible training datasets like the Pile or Common Crawl, both of which contain copyrighted materials scraped from the internet,[16] and which have been utilized by Anthropic in its training data.[17] Other datasets comprised of public domain data are also freely available for AI developers to use, such as Common Corpus,[18] which contains 500 billion words, roughly the same dataset size as that used to train OpenAI's GPT-3 model.[19]

21. AI companies may also procure "synthetic" datasets or generate synthetic datasets themselves. Synthetic datasets are datasets consisting of material that has itself been generated by a GAI model. AI companies can obtain pre-compiled synthetic datasets from third parties, such as

---

[14] Annie Gilbertson and Alex Reisner, *Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI*, PROOF (Jul. 16, 2024), https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai/.

[15] *See* Annie Gilbertson & Alex Reisner, *Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI*, Proof News (Jul. 16, 2024), https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai/; Declaration of Timothy Chung ¶¶ 9, 23, Exs. G, U.

[16] *See* Annie Gilbertson and Alex Reisner, *Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI*, PROOF (Jul. 16, 2024), https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai/.; Complaint, *The New York Times Company v. Microsoft Corp., et al.*, No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023), ECF No. 1 at 26–28.

[17] Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment* 8, 27, ANTHROPIC (Dec. 9, 2021), https://arxiv.org/pdf/2112.00861.pdf; Annie Gilbertson and Alex Reisner, *Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI*, PROOF (Jul. 16, 2024), https://www.wired.com/story/youtube-training-data-apple-nvidia-anthropic/.

[18] Pierre-Carl Langlais, *Releasing Common Corpus: the largest public domain dataset for training LLMs*, HUGGING FACE (Mar. 20, 2024), https://huggingface.co/blog/Pclanglais/common-corpus.

[19] *Id.*; Kate Knibbs, *Here's Proof You Can Train an AI Model Without Slurping Copyrighted Content*, WIRED (Mar. 20, 2024), https://www.wired.com/story/proof-you-can-train-ai-without-slurping-copyrighted-content/ (last visited Jul. 28, 2024).

the Cosmopedia dataset[20]; or they can generate synthetic datasets themselves, using their own or third-party GAI models.

22. Finally, AI companies may procure data directly from data owners or copyright holders, often by obtaining licenses for discrete libraries of content. In most instances, this method ensures AI companies obtain the proper authorization from the copyright holders to exploit their works and properly renumerate them for their use.

IV. Rights-Cleared Content and Authorized Uses

A. Justifications for Seeking Authorization

23. Although many AI companies do not seek authorization from copyright owners before exploiting their content for training materials, with many asserting fair use arguments to justify their unauthorized uses,[21] in my experience there are myriad reasons why AI companies choose to or should choose to seek authorization for the content they use to train their models.

24. First and foremost, respecting the rights of copyright holders with respect to training content is ethically sound since it allows authors and creators, such as writers and musicians, to approve and receive compensation for the use of their works. It also allows creators to have at least some say in how their works are exploited, preventing objectionable or harmful uses by AI on a case-by-case basis.

25. Second, using rights-cleared content is often a prudent business decision. Large copyright holders are among potential customers of AI products, and the unauthorized exploitation of their works is likely to alienate them, which may result in AI companies losing potential market share to their competitors. Relatedly, representations by developers that their AI products are built using rights-cleared content may serve as a draw to customers who are attracted by the ethical implications of such a practice or cannot otherwise practicably work with companies that exploit copyrighted content without authorization.

---

[20] *Cosmopedia v.0.*1, HUGGING Face, https://huggingface.co/datasets/HuggingFaceTB/cosmopedia (last visited Jul. 28, 2024).

[21] *See, e.g., Public Comments of Anthropic PBC*, ANTHROPIC (Oct. 30, 2023), at 7–9, https://perma.cc/8DBX-H3XU.

26. Third, seeking authorization prior to exploitation can guard against potential legal liability. As evidenced by the many lawsuits filed by various publishers and copyright holders against companies like OpenAI and Meta, unauthorized exploitation of copyrighted content for training purposes can invite lawsuits and challenges to those uses.

27. Fourth, a practice of using rights-cleared content can help AI companies in the hiring market. Some AI researchers, engineers, and other existing and potential employees of AI companies (myself included) have significant reservations about the ethical implications of the unauthorized exploitation of copyrighted content for training purposes. Seeking authorization for copyrighted materials in training datasets may help AI companies attract and retain critical talent in a competitive market.

28. Finally, obtaining authorization from copyright holders allows AI developers to directly source well-organized, cleaned-up data of the highest quality. This is particularly the case where licenses are obtained from large, institutional copyright holders with vast libraries of pre-vetted content, such as newspapers or music publishers.

B. Feasibility

29. While a given LLM trained solely on licensed and public domain data may not match the performance at the same cost of a model that uses copyrighted content without authorization (since the latter model would likely have easier access to more training data at a lower cost), I believe it is technologically and economically feasible to develop useful LLMs that are trained solely on a combination of licensed and public domain content.

30. Today, a number of GAI developers have taken the approach of training their models solely on rights-cleared or public domain data, including all models certified under the Licensed Model certification scheme from Fairly Trained.[22] For example, 273 Ventures' LLM,

[22] *Licensed Model Certification*, FAIRLY TRAINED, https://www.fairlytrained.org/certifications (last visited Jul. 26, 2024).

dubbed KL3M, was trained on datasets wholly comprised of public domain documents and explicitly licensed content, including legal, financial, and regulatory documents.[23]

31. As explained by Dr. Michael Smith, paying for copyrighted data is also economically feasible thanks to the rapid development of an AI training content licensing market, with major AI companies entering into multi-million dollar deals with copyright holders and dataset licensing services now offering catalogs of copyrighted content for AI training.[24]

32. Upfront costs for licensing can also be greatly reduced or deferred using alternative fee arrangements. For example, AI startup Bria AI entered into partnerships with Getty and 20 other stock image companies, who provide their images as training data on a revenue-sharing basis.[25] Stable Audio, which my team at Stability AI built, was trained on music provided by a stock audio library called AudioSparx, in a deal that covered revenue sharing between the two companies;[26] and Shutterstock licenses its library to AI developers, paying artist contributors based on how much an AI developer pays Shutterstock.[27]

33. Online marketplaces that simplify payment and licensing are also well on the way. Human Native AI will soon allow AI developers to easily source high quality data from copyright holders, such as publishers, record labels, film studios, and more, through its comprehensive online AI data marketplace, compensating copyright holders through recurring subscriptions or revenue share arrangements.[28] It already has access to text, video, images, and audio data, which it is able

---

[23] KL3M, https://www.kl3m.ai/ (last visited Jul. 28, 2024); Kate Knibbs, *Here's Proof You Can Train an AI Model Without Slurping Copyrighted Content*, WIRED (Mar. 20, 2024), https://www.wired.com/story/proof-you-can-train-ai-without-slurping-copyrighted-content/.

[24] Declaration of Michael Smith ("Smith Decl."), ¶¶ 76–81.

[25] Nico Grant and Cade Metz, *The Push to Develop Generative A.I. Without All the Lawsuits*, N.Y. TIMES (Jul. 19, 2024), https://www.nytimes.com/2024/07/19/technology/generative-ai-getty-shutterstock.html.

[26] Kyle Wiggers, *Stability AI, gunning for a hit, launches an AI-powered music generator*, TECHCRUNCH (Sep. 13, 2023), https://techcrunch.com/2023/09/13/stability-ai-gunning-for-a-hit-launches-an-ai-powered-music-generator/.

[27] Nico Grant and Cade Metz, *The Push to Develop Generative A.I. Without All the Lawsuits*, N.Y. TIMES (Jul. 19, 2024), https://www.nytimes.com/2024/07/19/technology/generative-ai-getty-shutterstock.html.

[28] *AI Data Marketplace*, HUMAN NATIVE AI, https://www.humannative.ai/ (last visited Jul. 26, 2024).

to license to AI companies for training through its limited access beta.[29] LANDR offers similar opportunities to musicians through its online Fair Trade AI program, distributing payment to musicians who allow their content to be used for AI training purposes.[30]

34. Datasets sourced with the proper authorizations can be further expanded using public domain datasets such as Common Corpus.

35. Given all these developments, while there may be challenges to using exclusively rights-cleared content in certain circumstances, that is no reason for making no effort to license at all.

V. Impact on Authors, Creators, and Copyright Holders

36. The failure to seek authorization for the use of copyrighted content as AI training data harms authors, creators, and copyright holders in a variety of ways. As Dr. Michael Smith explains, the economic harms are many, ranging from a usurpation of the authors and creators' markets, to the diversion of royalties, and the erosion of the incentive to create.[31]

37. Equally important are ethical considerations—that failing to credit or compensate authors, creators, and copyright holders robs them of recognition and reward for expending their creative efforts and care to create works of art, music, literature, and more.

38. Furthermore, and critically, unauthorized wholesale use of copyrighted content robs authors and creators of control over how and where their works are used. Some object to the use of generative AI in the creative process; others disapprove of the way certain AI companies treat creators and their work; still others are opposed to other harms caused by GAI, such as deepfakes and environmental harms. If the status quo continues, authors and creators will not be able to effectively control how their work is used in applications they do not want to support, especially for uses that are not fair under copyright law. This is particularly true given that the most widely used method of opting out of GAI training, updating a website's robots.txt file, is not guaranteed

[29] *See Sign Up*, HUMAN NATIVE AI, https://www.humannative.ai/sign-up (last visited Jul. 26, 2024).
[30] *LANDR Fair Trade AI Program*, LANDR, https://www.landr.com/fairai/ (last visited Jul. 26, 2024).
[31] Smith Decl. ¶¶ 46–61.

to be effective, for many reasons. First, such measures rely on voluntary compliance by GAI companies. Second, AI companies, including Anthropic, sometimes replace their web scrapers with new ones, resulting in websites blocking the wrong scrapers.[32]

39. GAI undoubtedly has many positive use cases and the potential to enhance the creative abilities of authors and creators. LLMs can be used by authors to assist in all manner of writing and creative tasks. I applaud the ability of GAI models to perform all of these tasks. But this functionality should not be built on widespread unauthorized exploitation of copyrighted materials at the expense of copyright holders.

I declare under penalty of perjury under the laws of the United States that the foregoing is true and correct to the best of my personal knowledge and belief.

Ed Newton-Rex

Executed in St. Helena, CA this 1st day of August, 2024

[32] Jason Koebler, *Websites are Blocking the Wrong AI Scrapers (Because AI Companies Keep Making New Ones)*, 404 (Jul. 29, 2024), https://www.404media.co/websites-are-blocking-the-wrong-ai-scrapers-because-ai-companies-keep-making-new-ones/.

SIGNATURE ATTESTATION PURSUANT TO CIVIL L.R. 5-1(h)

Pursuant to Civil L.R. 5-1(h), 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: August 1, 2024

/s/ Timothy Chung
Timothy Chung