

1 Joseph R. Saveri (State Bar No. 130064)
2 **JOSEPH SAVERI LAW FIRM, LLP**
3 601 California Street, Suite 1505
4 San Francisco, CA 94108
5 Telephone: (415) 500-6800
6 Facsimile: (415) 395-9940
7 Email: jsaveri@saverilawfirm.com

8 Matthew Butterick (State Bar No. 250953)
9 1920 Hillhurst Avenue, #406
10 Los Angeles, CA 90027
11 Telephone: (323) 968-2632
12 Facsimile: (415) 395-9940
13 Email: mb@buttericklaw.com

14 Laura M. Matson (*pro hac vice* pending)
15 **LOCKRIDGE GRINDAL NAUEN PLLP**
16 100 Washington Avenue South, Suite 2200
17 Minneapolis, MN 55401
18 Telephone: (612) 339-6900
19 Facsimile: (612) 339-0981
20 Email: lmmatson@locklaw.com

21 *Counsel for Individual and Representative*
22 *Plaintiffs and the Proposed Class*
23 *(continues on signature page)*

24 **UNITED STATES DISTRICT COURT**
25 **NORTHERN DISTRICT OF CALIFORNIA**
26 **SAN FRANCISCO DIVISION**

27 **Jingna Zhang**, an individual;
28 **Sarah Andersen**, an individual;
Hope Larson, an individual; and
Jessica Fink, an individual;

Individual and Representative Plaintiffs,

v.

Google LLC, a Delaware limited liability company; and
Alphabet Inc., a Delaware corporation;

Defendants.

Case No.

COMPLAINT

CLASS ACTION

DEMAND FOR JURY TRIAL

1 Plaintiffs Jingna Zhang, Sarah Andersen, Hope Larson, and Jessica Fink (together “Plaintiffs”),
2 on behalf of themselves and all others similarly situated, bring this class-action complaint
3 (“Complaint”) against defendants Google LLC (“Google”) and Alphabet Inc. (“Alphabet”) (together
4 “Defendants”).

6 OVERVIEW

7 1. *Artificial intelligence*—commonly abbreviated “AI”—denotes software that is designed
8 to algorithmically simulate human reasoning or inference, often using statistical methods.

9 2. Imagen is an AI software product created, maintained, and sold by Google. Imagen is a
10 *text-to-image diffusion model*. A text-to-image diffusion model takes as input a short text description of
11 an image (also known as a *text prompt*) and then uses a machine-learning technique called *diffusion* to
12 generate an image in response to the prompt.

13 3. Rather than being programmed in the traditional way—that is, by human programmers
14 writing code—a diffusion model is *trained* by copying an enormous quantity of digital images with
15 associated text captions, extracting protected expression from these works, and transforming that
16 protected expression into a large set of numbers called *weights* that are stored within the model. These
17 weights are entirely and uniquely derived from the protected expression in the training dataset.
18 Whenever a diffusion model generates an image in response to a user prompt, it is performing a
19 computation that relies on these stored weights, with the goal of imitating the protected expression
20 ingested from the training dataset.

21 4. Training a model first requires amassing a huge corpus of data, called a *dataset*. The AI
22 models at issue in this complaint were trained on datasets containing millions of images paired with
23 descriptive captions. In this complaint, each image–caption pair is called a *training image*. During
24 training of the model, the training images in the dataset are directly copied in full and then completely
25 ingested by the model, meaning that protected expression from every training image enters the model.
26 As it copies and ingests billions of training images, the model progressively develops the ability to
27 generate outputs that mimic the protected expression copied from the dataset.

1 13. Plaintiff Hope Larson is a cartoonist and illustrator who lives in North Carolina.

2 14. Plaintiff Jessica Fink is a cartoonist and illustrator who lives in New York.

3 15. A nonexhaustive list of registered copyrights owned by Plaintiffs is included as
4 **Exhibit A: Plaintiff Copyright Registrations**. A nonexhaustive list of copyrighted images registered
5 by Plaintiffs and infringed by Defendants is included as **Exhibit B: Plaintiff Images in LAION-400M**.

6 16. The images shown in Exhibit B are offered as a representative sample of works by
7 Plaintiffs that appear in the LAION-400M dataset—not an exhaustive or complete list. Plaintiffs
8 confirmed that these particular images were in the LAION-400M dataset by searching for their own
9 names on two websites that allow searching of the LAION datasets: <https://haveibeenentrained.com> and
10 <https://rom1504.github.io/clip-retrieval/>. On information and belief, all of Plaintiffs' works that were
11 registered as part of the collections in Exhibit A and were online were scraped into the LAION-400M
12 dataset.

13
14 **DEFENDANTS**

15 17. Defendant Google LLC is a Delaware limited liability company with its principal place
16 of business at 1600 Amphitheatre Parkway, Mountain View CA 94043.

17 18. Defendant Alphabet Inc. is a Delaware corporation with its principal place of business at
18 1600 Amphitheatre Parkway, Mountain View CA 94043. In 2015, Google became a subsidiary of
19 Alphabet.

20
21 **AGENTS AND CO-CONSPIRATORS**

22 19. The unlawful acts alleged against the Defendants in this Complaint were authorized,
23 ordered, or performed by the Defendants' respective officers, agents, employees, representatives, or
24 shareholders while actively engaged in the management, direction, or control of the Defendants'
25 businesses or affairs. The Defendants' agents operated under the explicit and apparent authority of
26 their principals. Each Defendant, and its subsidiaries, affiliates, and agents operated as a single unified
27 entity.
28

1 26. Despite its professed commitment to “not release Imagen for public use without further
2 safeguards,”⁴ Google soon reversed course.

3 27. In November 2022, Google made Imagen publicly available to a select group of users
4 through its AI Test Kitchen app. According to reporting at the time, Google “announced it will be
5 adding Imagen—in a *very* limited form—to its AI Test Kitchen app as a way to collect early feedback on
6 the technology.”⁵

7 28. In January 2023, plaintiff Sarah Andersen and two other artists filed the first lawsuit in
8 the U.S. challenging the legality of training text-to-image diffusion models on copyrighted work without
9 consent, credit, or compensation. That case, *Andersen v. Stability AI et al.*, (Case No. 23-cv-00201,
10 N.D. Cal.) challenged two models similar to Imagen—called Stable Diffusion and Midjourney—both
11 of which were also trained on the LAION dataset. (The *Andersen* case is currently proceeding.)

12 29. In May 2023, Google made Imagen even more widely available through its commercial
13 AI cloud-computing service, called Vertex AI. According to a Google blog post about Vertex AI,
14 Google described it as “Imagen, our text-to-image foundation model, lets organizations generate and
15 customize studio-grade images at scale for any business need.”⁶

16 30. In October 2023, Google made Imagen even more widely available through a tool called
17 Search Generative Experience. According to reporting at the time, “If you’re opted in to [Search
18 Generative Experience] through Google’s Search Labs program, you can just type your query into the
19 Google search bar. After you do, [Search Generative Experience] can create a few images based on your
20 prompt that you can pick from. The tool is powered by the Imagen family of AI models.”⁷

21 31. In December 2023, Google released the successor to Imagen, called Imagen 2. Unlike
22 the paper that accompanied the initial version of Imagen, Google’s introduction of Imagen 2 carefully
23

24 ⁴ *Id.*

25 ⁵ See <https://www.theverge.com/2022/11/2/23434361/google-text-to-image-ai-model-imagen-test-kitchen-app>

26 ⁶ See <https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-launches-new-ai-models-opens-generative-ai-studio>

27 ⁷ See <https://www.theverge.com/2023/10/12/23913337/google-ai-powered-search-sge-images-written-drafts>
28

1 omits a detailed description of its training dataset. Google limits itself to vague comments such as
2 “From the outset, we invested in training data safety for Imagen 2, and added technical guardrails to
3 limit problematic outputs like violent, offensive, or sexually explicit content.”⁸

4 32. On information and belief, Google did not disclose details about the training dataset for
5 Imagen 2 because it was aware of the *Andersen v. Stability AI et al.* case and hoped to avoid being named
6 as a defendant in a lawsuit over the legality of training on mass quantities of copyrighted works without
7 consent, credit, or compensation.

8 33. On information and belief, Google included LAION-400M in its training dataset for
9 Imagen 2, because a) it had already done so for the first version of Imagen, and b) one of the architects
10 of the LAION image datasets, Romain Beaumont, is a Google employee, who Google hired specifically
11 to exercise influence over the LAION organization and its image datasets.

12 13 **A KEY SOURCE OF GOOGLE’S TRAINING DATA: LAION**

14 34. LAION (acronym for “Large-Scale Artificial Intelligence Open Network”) is an
15 organization based in Hamburg, Germany. According to its website, LAION is led by Christoph
16 Schuhmann. LAION’s stated goal is “to make large-scale machine learning models, datasets and
17 related code available to the general public.”⁹ All of LAION’s projects are made available for free.

18 35. Since 2021, a key member of LAION’s team has been Romain Beaumont, who describes
19 himself on the LAION website as an “open source contributor ... I like to apply scale and deep learning
20 to build AI apps and models.”¹⁰

21 36. LAION’s most well-known projects are the datasets of training images it has released
22 for training machine-learning models, which are now widely used in the AI industry.

23 37. In August 2021, LAION released LAION-400M, a dataset of 400 million training
24 images assembled from images accessible on the public internet. At the time, LAION-400M was the
25 largest freely available dataset of its kind. Until December 2023, LAION distributed the LAION-400M

26
27 ⁸ See <https://deepmind.google/technologies/imagen-2/>

28 ⁹ <https://laion.ai/about/>

¹⁰ See <https://laion.ai/team/>

1 dataset to the public through its own website and elsewhere. (In December 2023, due to the discovery
2 of child sexual-abuse material (“CSAM”) in the LAION datasets, the LAION organization retracted
3 these datasets—including LAION-400M—from the public internet.)

4 38. Also in August 2021, Romain Beaumont created an online tool called Clip Retrieval that
5 acted as a search interface to LAION to check whether certain artists or artworks were included in the
6 LAION-400M dataset.¹¹ Beaumont’s tool was popular. It was online until December 2023. (In
7 December 2023, it was disabled due to the aforementioned issues with CSAM in the LAION datasets.)

8 39. In November 2021, Romain Beaumont was a primary author of the paper that
9 introduced the LAION-400M dataset, titled “LAION-400M: Open Dataset of CLIP-Filtered 400
10 Million Image-Text Pairs,” released in November 2021 (hereafter, the “Beaumont-LAION Paper”).¹²

11 40. When one downloads the LAION-400M dataset, one gets a list of metadata records,
12 one for each training image. Each record includes the URL of the image, the image caption, a
13 measurement of the similarity of the caption and image, a NSFW flag (indicating the probability the
14 image contains so-called “not safe for work” content), and the width and height of the image.

15 41. The actual images referenced in the LAION-400M dataset records are not included
16 with the dataset. Anyone who wishes to use LAION-400M for training their own machine-learning
17 model must first acquire copies of the actual images from their URLs. To facilitate the copying of these
18 images, Romain Beaumont created a software tool called `img2dataset` that takes the LAION-400M
19 metadata records as input and makes copies of the referenced images from the URLs in each metadata
20 record, thereby creating local copies. The `img2dataset` tool is distributed from a page Beaumont
21 controls on GitHub.¹³ LAION promotes the `img2dataset` tool in its documentation for LAION-
22 400M. (“This metadata dataset purpose is to download the images for the whole dataset or a subset of
23 it by supplying it to the very efficient `img2dataset` tool.”¹⁴)

24
25
26 ¹¹ See <https://rom1504.github.io/clip-retrieval>

27 ¹² <https://arxiv.org/abs/2111.02114>

28 ¹³ <https://github.com/rom1504/img2dataset>

¹⁴ See <https://laion.ai/blog/laion-400-open-dataset/>

1 42. Training a model with the LAION-400M dataset cannot begin without first using
2 `img2dataset` or another similar tool to download the images in the dataset. Thus, because Google has
3 trained Imagen on LAION-400M, Google has necessarily made one or more copies of images
4 belonging to Plaintiffs as shown in Exhibit B, either by using Romain Beaumont’s `img2dataset` tool or
5 another. Plaintiffs never authorized any of these LAION dataset users to copy their images or use them
6 for training any models.

7 43. One of the entities that has made unauthorized copies of the LAION-400M training
8 images is LAION itself. According to the Beaumont–LAION Paper, LAION made the dataset by
9 starting with Common Crawl metadata records. Common Crawl is a corpus of 250 billion web pages
10 copied from the public web, including assets like Plaintiffs’ images (<https://commoncrawl.org/>). The
11 metadata records contain web URLs. According to the Beaumont–LAION Paper, LAION created
12 training images by first “pars[ing] through [the metadata records] from Common Crawl and pars[ing]
13 out all HTML IMG tags containing an alt-text attribute [that is, a text caption].” Then, LAION
14 “download[ed] the raw images from the parsed URLs”. Beaumont–LAION Paper at 3.

15 44. Sometime after the release of LAION-400M in August 2021, a company called
16 Stability AI funded LAION’s creation of a similar dataset, but much larger. In March 2022, Stability AI
17 CEO Mostaque called himself “the biggest backer of LAION.”¹⁵

18 45. But Google wasn’t far behind. In March 2022, Google hired Romain Beaumont as a full-
19 time software engineer, a position he has held since. On information and belief, Google hired Beaumont
20 primarily to influence the creation of future LAION image datasets, based on a) Beaumont’s key role
21 creating LAION-400M—which Google used to train Imagen; b) Beaumont’s control of the
22 `img2dataset` tool that was essential to using the LAION-400M dataset, and c) Beaumont’s control of
23 the Clip Retrieval website that was essential to searching the LAION-400M dataset.

24 46. Later in March 2022, LAION released LAION-5B, a dataset of 5.85 billion training
25 images—more than 14 times bigger than LAION-400M. The author of the LAION blog post
26 announcing LAION-5B was Romain Beaumont.¹⁶

27 _____
28 ¹⁵ <https://discord.com/channels/662267976984297473/938713143759216720/954674533942591510>

¹⁶ See <https://laion.ai/blog/laion-5b/>

1 47. In August 2022, Romain Beaumont created a specialized AI model to rate the aesthetic
2 quality of an image, and used this model to create subsets of the LAION-5B training images filtered by
3 aesthetic quality, which Beaumont called LAION-Aesthetics. In its introduction of Imagen 2 in
4 December 2023, Google said “We trained a specialized image aesthetics model based on human
5 preferences for qualities like good lighting, framing, exposure, sharpness, and more. Each image was
6 given an aesthetics score which helped condition Imagen 2 to give more weight to images in its training
7 dataset that align with qualities humans prefer.”¹⁷ On information and belief, Beaumont’s work on
8 LAION-Aesthetics formed the basis of Imagen 2’s “aesthetics model”, since at the time Beaumont was
9 both a contributor to LAION and a full-time employee of Google.

10 48. In October 2022, Romain Beaumont was a primary author of the paper about LAION-
11 5B, called “LAION-5B: An open large-scale dataset for training next generation image-text models.”
12 (hereafter, the “Beaumont-LAION-5B Paper”). According to the Beaumont-LAION-5B Paper,
13 LAION-400M is a subset of LAION-5B, meaning every image in LAION-400M is also in LAION-5B.

14 49. Just like the LAION-400M dataset, the actual images referenced in the LAION-5B
15 dataset records are not included with the dataset. Anyone who wishes to use LAION-5B for training
16 their own machine-learning model must first acquire copies of the actual images from their URLs. As
17 mentioned above, to facilitate the copying of these images, Romain Beaumont created a software tool
18 called `img2dataset` that takes the LAION-5B metadata records as input and makes copies of the
19 referenced images from the URLs in each metadata record, thereby creating local copies. The
20 `img2dataset` tool is distributed from a page Beaumont controls on GitHub.¹⁸

21
22
23
24
25
26
27

¹⁷ See <https://deepmind.google/technologies/imagen-2/>

28 ¹⁸ <https://github.com/rom1504/img2dataset>

COUNT 1

DIRECT COPYRIGHT INFRINGEMENT (17 U.S.C. § 501)

AGAINST GOOGLE

50. The preceding factual allegations are incorporated by reference.

51. As the owners of the registered copyrights in the works in Exhibit B, Plaintiffs hold the exclusive rights to those works under the U.S. Copyright Act (17 U.S.C. § 106).

52. Plaintiffs never authorized Google to use their copyrighted work in any way. Nevertheless, Google repeatedly violated Plaintiffs' exclusive rights under § 106 and continues to do so today. Plaintiffs and the Class members never authorized Google to make copies of their works, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works).

53. On information and belief, Google has used Plaintiffs' training images to train other versions of Imagen, including Imagen 2, and so-called "multimodal" models that are trained on training images as well as text, such as Google Gemini. Collectively, Imagen and other models that Google trained on LAION-400M are called the **Google-LAION Models**.

54. The LAION-400M dataset contains only URLs of training images, not the actual training images. Therefore, anyone who wishes to use LAION-400M for training their own machine-learning model must first acquire copies of the actual training images from their URLs. Consistent with this, in preparation for training the Google-LAION Models, Google made one or more copies of the LAION-400M training images, including the Plaintiff works in Exhibit B, so they could be fed to the Google-LAION Models as training data. The copies made of each copyrighted work were substantially similar to that copyrighted work.

55. During the training of the Google-LAION Models, Google made a series of intermediate copies of the LAION-400M training images, including the Plaintiff works in Exhibit B. The intermediate copies of each copyrighted work that Google made during training of the Google-LAION Models were substantially similar to that copyrighted work.

1 63. **Class definition.** Plaintiffs bring this action for damages and injunctive relief as a class
2 action under Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3), on behalf of the following
3 Class:

4 **All persons or entities domiciled in the United States that own a**
5 **United States copyright in any work that Google used as a training**
6 **image for the Google-LAION Models during the Class Period.**

7 64. This Class definition excludes:

- 8 a. Defendants named herein;
- 9 b. any of the Defendants' co-conspirators;
- 10 c. any of Defendants' parent companies, subsidiaries, and affiliates;
- 11 d. any of Defendants' officers, directors, management, employees, subsidiaries,
12 affiliates, or agents;
- 13 e. all governmental entities; and
- 14 f. the judges and chambers staff in this case, as well as any members of their
15 immediate families.

16 65. **Numerosity.** Plaintiffs do not know the exact number of members in the Class. This
17 information is in the exclusive control of Defendant. On information and belief, there are at least
18 thousands of members in the Class geographically dispersed throughout the United States. Therefore,
19 joinder of all members of the Class in the prosecution of this action is impracticable.

20 66. **Typicality.** Plaintiffs' claims are typical of the claims of other members of the Class
21 because Plaintiffs and all members of the Class were damaged by the same wrongful conduct of
22 Defendant as alleged herein, and the relief sought herein is common to all members of the Class.

23 67. **Adequacy.** Plaintiffs will fairly and adequately represent the interests of the members of
24 the Class because the Plaintiffs have experienced the same harms as the members of the Class and have
25 no conflicts with any other members of the Class. Furthermore, Plaintiffs have retained sophisticated
26 and competent counsel who are experienced in prosecuting federal and state class actions, as well as
27 other complex litigation.
28

- 1 e) Pre- and post-judgment interest on the damages awarded to Plaintiffs and the Class, and
2 that such interest be awarded at the highest legal rate from and after the date this class
3 action complaint is first served on Defendant.
- 4 f) Defendants are to be jointly and severally responsible financially for the costs and
5 expenses of a Court approved notice program through post and media designed to give
6 immediate notification to the Class.
- 7 g) Further relief for Plaintiffs and the Class as may be just and proper.
- 8

9 **JURY TRIAL DEMANDED**

10 Under Federal Rule of Civil Procedure 38(b), Plaintiffs demand a trial by jury of all the claims
11 asserted in this Complaint so triable.

12

13 Dated: April 26, 2024

14 By: /s/ Joseph R. Saveri
Joseph R. Saveri

15 Joseph R. Saveri (State Bar No. 130064)
16 Cadio Zirpoli (State Bar No. 179108)
17 Christopher K. L. Young (State Bar No. 318371)
18 Elissa Buchanan (State Bar No. 249996)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
eabuchanan@saverilawfirm.com

23 Matthew Butterick (State Bar No. 250953)
24 1920 Hillhurst Avenue, #406
25 Los Angeles, CA 90027
26 Telephone: (323) 968-2632
27 Facsimile: (415) 395-9940
28 Email: mb@buttericklaw.com

1 Brian D. Clark (*pro hac vice* pending)
2 Laura M. Matson (*pro hac vice* pending)
3 Arielle S. Wagner (*pro hac vice* pending)
4 Eura Chang (*pro hac vice* pending)
5 **LOCKRIDGE GRINDAL NAUEN PLLP**
6 100 Washington Avenue South, Suite 2200
7 Minneapolis, MN 55401
8 Telephone: (612) 339-6900
9 Facsimile: (612) 339-0981
10 Email: bdclark@locklaw.com
11 lmmatson@locklaw.com
12 aswagner@locklaw.com
13 echang@locklaw.com

14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
*Counsel for Individual and Representative
Plaintiffs and the Proposed Class*