William Dreher (*pro hac vice*)
Derek W. Loeser (*pro hac vice*)
Cari C. Laufenberg (*pro hac vice*)
Benjamin Gould (SBN 250630)
KELLER ROHRBACK L.L.P.
1201 Third Avenue, Suite 3400
Seattle, WA 98101-3268
(206) 623-1900
wdreher@kellerrohrback.com
dloeser@kellerrohrback.com
claufenberg@kellerrohrback.com
bgould@kellerrohrback.com

Matthew Butterick (SBN 250953)
BUTTERICK LAW PC
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
(323) 968-2632
mb@buttericklaw.com

Rohit D. Nath (SBN 316062)
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
(310) 789-3100
RNath@susmangodfrey.com

Justin A. Nelson (*pro hac vice*)
Alejandra C. Salinas (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
(713) 653-7802
JNelson@susmangodfrey.com
ASalinas@susmangodfrey.com

Additional Counsel Listed in Signature Block

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
OAKLAND DIVISION

GRADY HENDRIX ET AL.,

Individual and Representative Plaintiffs,

v.

APPLE INC.,

Defendant.

This Document Relates To:
All Consolidated Actions

No. 4:25-cv-07558-YGR

FIRST AMENDED CONSOLIDATED CLASS ACTION COMPLAINT

FILED UNDER SEAL

JURY TRIAL DEMANDED

Plaintiffs Grady Hendrix, Jennifer Roberson, John Hornor Jacobs, and Eboni Mckinnon Smith (a/k/a Kianna Alexander) (together "Plaintiffs"), on behalf of themselves and all others similarly situated (the "Class," as defined below), bring this class-action complaint ("Complaint") against Defendant Apple Inc. ("Apple" or "Defendant").

I. INTRODUCTION

1. Apple, one of the world's most valuable companies, has invested substantial capital and engineering resources into developing generative artificial intelligence ("AI") programs and technologies. Among Apple's AI products is Apple Intelligence. Apple regards Apple Intelligence as a breakthrough innovation that will make its users' experiences "profoundly different" across various product applications. Through Apple Intelligence, Apple hopes to add trillions to its market capitalization in coming years.

2. But Apple is building this new enterprise in part by using Books3, a dataset of pirated copyrighted books that includes the published works of Plaintiffs and the Class. Apple has developed, designed, maintained, and commercialized multiple large language models ("LLMs"). An LLM is a type of machine-learning model designed to emit convincingly naturalistic text outputs in response to user prompts. Companies developing LLMs "train" the LLMs to produce these outputs by feeding the models copies of an enormous quantity of data—frequently including massive numbers of textual works. Apple used the dataset called Books3 to train one set of LLMs, the OpenELM language models. Apple also, on information and belief, trained a second set of LLMs, its Foundation Language Models, using this same pirated dataset.

3. Apple's generative AI also uses Applebot, a software program that copies mass quantities of webpages (also known as "scraping"). Apple scraped data with Applebot for nearly nine years before disclosing that it intended to train its commercial AI products on this scraped data. Scrapers like Applebot can also reach so-called "shadow libraries" that host millions of other unlicensed copyrighted books, including, on information and belief, Plaintiffs' and Class Members' copyrighted works. Shadow libraries are the approximate equivalent of Napster for books; users upload unauthorized copies of books, and the shadow library helps other users download and access the works for free without paying the customary price. As has become public in other cases, a number of Apple's competitors in the AI race—including OpenAI, Meta Platforms, Anthropic, and NVIDIA—resorted to obtaining copyrighted materials from shadow libraries like Library Genesis, Z-Library, Anna's Archive, Sci-Hub, and Pirate Library Mirror. Upon information and belief, like its competitors, Apple

obtained works from these notorious pirate libraries—accessed either through Applebot or torrenting—to build its own permanent library of stolen material for training its models, for commercial research, and for other purposes.

4. Apple also illicitly obtained copyrighted books for AI training by other means, including by scanning books. After acquiring those books, Apple made additional copies of them. In some cases, Apple converted the books to new formats *and* made additional copies—for instance, when converting text to audio, and vice versa, and when generating synthetic data (that is, data created by a generative AI model)—from models trained on books. And in all of these instances, Apple stored such data in central repositories and made such data available for additional future uses, including to make training more convenient.

5. The quality of Apple's models depends on the data on which it is trained. LLMs operate by copying and later simulating creative expression found in copyrighted works. For this reason, the inclusion of expressive, high-quality material—especially copyrighted material—in Apple's AI training datasets is deliberate and commercially significant. For instance, to access even more copyrighted material to develop its valuable generative AI products, Apple entered into a multimillion-dollar licensing agreement with Shutterstock. But not with Plaintiffs or the Class.

6. Plaintiffs and the Class are owners of registered copyrights for their published works. They did not consent to the use of their works in any model developed by Apple, including the Foundation Language Models and OpenELM language models.

7. The licensing market for AI training data is robust and growing. For instance, in February 2026, Microsoft launched a data-licensing marketplace called the Publisher Content Marketplace; Amazon has held discussions with publishers about building a similar service. Nevertheless, Apple did not compensate creators for use of their copyrighted works and concealed the sources of the datasets they downloaded, curated, and used for AI training to evade legal scrutiny. On information and belief, Apple continues to retain its own curated, private shadow library of pirated books to train AI models currently or soon to be under development and for other commercial purposes, all without seeking Plaintiffs' or Class Members' consent or providing them compensation.

8. Apple has copied the copyrighted works of Plaintiffs and the Class to train AI models whose outputs compete with and dilute the market for those very works. Without those works, Apple's generative AI models, Apple products that integrate those models via Apple Intelligence, and Apple ecosystems in which third-party products integrate those models (such as the Apple App Store) would have far less commercial value. This conduct has deprived Plaintiffs and the Class of control over their work, undermined the economic value of their labor, and positioned Apple to achieve massive commercial success through unlawful means.

9. In sum, Apple engaged in at least the following eight infringing uses as to Plaintiffs and the Class, as detailed in the remainder of this Complaint:

A. *Acquiring Data via Pirated Sources:* Apple made unauthorized reproductions of Plaintiffs' works as part of Apple's downloading, torrenting, or other mode of initial acquisition of one or more datasets of pirated works, including at least Books3 and Library Genesis.

B. *Acquiring Data via Applebot:* Apple made unauthorized copies of Plaintiffs' works as part of Apple's scraping of the public internet using Applebot or other methods;

C. *Reproducing Data after Acquisition:* Apple made unauthorized copies of Plaintiffs' works whether for the purpose of training, research, transfer, or permanent retention, regardless of the mode of acquisition (including but not limited to downloading, scraping, licensing, contracting, and scanning);

D. *Curating:* Apple made unauthorized copies of Plaintiffs' works when filtering, cleaning, and processing at least one pirated dataset, such as Books3, and components of such datasets, to create a curated, private shadow library of datasets for use in various stages of training LLMs;

E. *Training OpenELM Models:* Apple made unauthorized copies of Plaintiffs' works when training its OpenELM models;

F. *Training FLMs:* Apple made unauthorized copies of Plaintiffs' works when training all versions and generations of its Foundation Language Models;

G. *Training Classifier and Teacher Models:* Apple made unauthorized copies of Plaintiffs' works when training classifier and teacher models; and

H. *Retaining for Private Data Library:* Apple engaged in unauthorized retention, curation, reproduction, and use of Plaintiffs' works along with all the data Apple has illegally downloaded, curated, reproduced, and used thus far, in the form of a private data library for use in models currently under development as well as future models—an AI data library that includes the Books3 and Library Genesis dataset, which, in turn, includes Plaintiffs' works.

II. JURISDICTION AND VENUE

10. This Court has subject-matter jurisdiction under 28 U.S.C. § 1331 because this case arises under the Copyright Act (17 U.S.C. § 501).

11. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Apple is headquartered in this district.

III. DIVISIONAL ASSIGNMENT

12. Under Civil Local Rule 3-2(c), assignment of this case to the Oakland Division is proper because this case pertains to intellectual-property rights, which is a district-wide case category under General Order No. 44, and therefore venue is proper in any courthouse in this District.

IV. PARTIES

A. Plaintiffs

13. Plaintiff Grady Hendrix is an author who lives in New York. He owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "Hendrix Asserted Works"): *Horrorstor* (U.S. Copyright Office Registration Number TX0007957438); *My Best Friend's Exorcism* (USCO Registration Number TX0008276604); *Paperbacks from Hell* (USCO Registration Number TX0008461565); *We Sold Our Souls* (USCO Registration Number TX0008645012); *Ankle Snatcher* (USCO Registration Number

TX0009327754); *BadAsstronauts* (USCO Registration Number TX0009281005); *How to Sell a Haunted House* (USCO Registration Number TX0009231440); *The Final Girl Support Group* (USCO Registration Number TX0009212410); *The Southern Book Club's Guide to Slaying Vampires: A Novel* (USCO Registration Number TX0008888592, TX0008986901); and *Witchcraft for Wayward Girls* (USCO Registration Number TX0009505624).

14. Plaintiff Jennifer Roberson is an author who lives in Arizona. She owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "Roberson Asserted Works"): *Life and Limb* (USCO Registration Number TX0008936965); *Shapechangers* (USCO Registration Number TX0001405603); *Legacy of the Sword* (USCO Registration Number TX0002124929); *Track of the White Wolf* (USCO Registration Number TX0002138857); *A Pride of Princes* (USCO Registration Number TX0002595154); *Daughter of the Lion* (USCO Registration Number TX0002655076); *Flight of the Raven* (USCO Registration Number TX0003038048); *A Tapestry of Lions* (USCO Registration Number TX0003445261); *Karavans* (USCO Registration Number TX0006494921); *Deepwood* (USCO Registration Number TX0006608590); *The Wild Road* (USCO Registration Number TX0007614823); *Lady of the Glen* (USCO Registration Number TX0004292057); *Lady of the Forest* (USCO Registration Number TX0003467465); *Lady of Sherwood* (USCO Registration Number TX0005700486); *Sword-Dancer* (USCO Registration Number TX0002125935); *Sword-Singer* (USCO Registration Number TX0002585643); *Sword-Maker* (USCO Registration Number TX0003031154); *Sword-Breaker* (USCO Registration Number TX0003187269); *Sword-Born* (USCO Registration Number TX0004752101); *Sword-Sworn* (USCO Registration Number TX0005504873); *Sword-Bound* (USCO Registration Number TX0007724532); *The Song of Homana* (USCO Registration Number TX0001840915); *Mad Jack* (USCO Registration Number TX0004840105); *Out of Avalon* (USCO Registration Number TX0005450939 & TX0005450940); *Sinners and Saints* (USCO Registration Number TX0009100077); *Smoketree* (USCO Registration Number TX0001570167); and *Sword-Bearer* (USCO Registration Number TX0009259320).

15. Plaintiff John Hornor Jacobs is an author who lives in Arkansas. He owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "Hornor Jacobs Asserted Works"): *The Conformity* (USCO Registration Number TX0008114163); *The Sea Dreams It Is the Sky* (USCO Registration Number TX0008830088); *The Shibboleth* (USCO Registration Number TX0007905434); *This Dark Earth* (USCO Registration Number TX0007561564); *Twelve-Fingered Boy* (USCO Registration Number TX0009572498); *A Lush and Seething Hell* (USCO Registration Number TX0008804502); and *My Heart Struck Sorrow* (USCO Registration Number TX0008809460).

16. Plaintiff Eboni Mckinnon Smith (a/k/a Kianna Alexander) is an author who lives in North Carolina. She owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "Mckinnon Smith Asserted Works"): *Love for All Time* (USCO Registration Number TX0008672549); *Couldn't Ask for More* (USCO Registration Number TX0008668191); *Tempo of Love* (USCO Registration Number TX0008448578) (re-released in 2023 as *Stardust*); *Back to Your Love* (USCO Registration Number TX0009170932); *This Tender Melody* (USCO Registration Number TX0008193076); *Every Beat of My Heart* (USCO Registration Number TX0008217057); *After Hours Redemption* (USCO Registration Number TX0009021729); *A Love Like This* (USCO Registration Number TX0008563355); *After Hours Attractions* (USCO Registration Number TX0009100320); *After Hours Temptations* (USCO Registration Number TX000922722); Then Came You (USCO Registration Number TX0008787155); *Carolina Built* (USCO Registration Number TX0009135026); *After Hours Agenda* (USCO Registration Number TX0009297307); *Love for All Time* (USCO Registration Number TX0008672549).

17. Below, the Hendrix Asserted Works, Roberson Asserted Works, Hornor Jacobs Asserted Works, and Mckinnon Smith Asserted Works are referred to collectively as the "Infringed Works." On information and belief, Apple may have also infringed other works by Plaintiffs. A complete list of all of the Plaintiffs' registered copyrighted works is attached as Exhibit A to this Complaint.

B. Defendant

18. Defendant Apple Inc. is a California corporation with its principal place of business at One Apple Park Way, Cupertino CA 95014.

C. Agents and co-conspirators

19. The unlawful acts alleged against Apple in this class action complaint were authorized, ordered, or performed by Apple's respective officers, agents, employees, representatives, or shareholders while actively engaged in the management, direction, or control of Apple's businesses or affairs. Apple's agents operated under the explicit and apparent authority of their principals. Apple and its subsidiaries, affiliates, and agents operated as a single unified entity.

20. Various persons or firms not named as defendants may have participated as co-conspirators in the violations alleged herein and may have performed acts and made statements in furtherance thereof. Each acted as the principal, agent, or joint venture of Apple with respect to the acts, violations, and common course of conduct alleged herein.

V. FACTUAL ALLEGATIONS

21. Apple, currently the second largest company in the world by market capitalization, is an electronics and media company that designs, manufactures, and sells software and hardware technology products. Every second, Apple sells seven iPhones. Apple owns and operates the App Store, the primary platform for app developers to offer, and users to download, purchase, and subscribe to, mobile applications that run on Apple's mobile operating system, iOS. The App Store earns billions of dollars of revenue for Apple each year.

22. In or around April 2024, Apple released a suite of open-source generative-AI models called OpenELM. Just a few months later, in or around June 2024, Apple announced the development of its commercial artificial-intelligence platform, called Apple Intelligence. Apple Intelligence includes multiple generative-AI models and related tools and technologies. The day after Apple officially introduced Apple Intelligence—what one reporter likened to ushering in a "new era"—the company gained more than $200 billion in value: "the single most lucrative day in the history of the company."

23. In January 2025, the company reported its "best quarter ever" with revenue of $124.3 billion, twice citing Apple Intelligence in its press release announcing the same and deeming it a part of the company's "best-ever lineup of products and services." The technology is integrated across Apple's products, including iPhones, and is intended to "make[] apps and experiences even better and more personal." Analysts contend that Apple Intelligence could add $4 trillion to the company's market capitalization.

24. To train its generative-AI models, including the models that are part of Apple Intelligence, Apple first amassed an enormous library of data. Part of Apple's data library includes copyrighted works—including books created by Plaintiffs—that were copied without author or owner consent, credit, or compensation. In essence, Apple has created its own internal shadow library for the purpose of training current and future LLMs that are and will be integrated into Apple software, Apple apps, and third-party apps that are offered on the Apple App Store and run on Apple hardware.

25. Apple has not attempted to pay these authors and owners for their contributions to this potentially lucrative venture. Apple did not seek the licenses necessary to copy and use the copyrighted books provided to its models. Instead, it intentionally evaded payment by multiple methods: using books already compiled in pirated datasets, scraping books off the web, contracting for books from third parties, and scanning and copying books. Nor did Apple pay these authors and owners to make copies of their works for other purposes, beyond training.

A. How large language models work.

26. Artificial intelligence denotes software that is designed to algorithmically create an illusion of human reasoning or inference, often using statistical and mathematical methods. While called "artificial intelligence," these systems largely derive their value from human intelligence, as embodied within (among other things) the massive number of copyrighted works that are used to train them.

27. As noted above, though an LLM is a software program, it is not created the way most software programs are—by human software programmers writing code. Rather, LLMs like Apple's are trained by copying an enormous quantity of textual works and then feeding these copies into the

model. Typically, an LLM developer like Apple starts by identifying and copying its initial raw datasets containing such works. Then, these raw datasets are filtered and cleaned before they are used as model training data. For example, model developers often create software filters to remove certain categories of data that are undesirable for training, such as personally identifying information, copyright and license notices, or profanity and unsafe material. During this process, LLM developers often create derivative datasets that are subsets or altered versions of the original raw datasets. Importantly, because these additional derivative datasets also contain copies of parts of or entire textual works from the original raw datasets, including copyrighted works, the process of filtering and cleaning raw datasets to create usable training data for an LLM necessarily requires the creation of additional copies of the textual works in the raw datasets. Furthermore, because this gathering and processing of data is considered a strategic investment, the model developer often retains the datasets in a private data library, to remain available for future research and reuse.

28. The corpus of text ultimately used to train a model is called the "training dataset." Datasets may have multiple uses during the development process of an LLM—for instance, checking the performance of a model at intermediate stages of training, in an iterative development process. Therefore, even if a dataset is not part of a model's final training dataset, it may still be used in a model's development process and retained in a data library for use in development of a future model.

29. During training, the LLM copies and ingests each textual work in the training dataset and extracts protected expression from it. The LLM progressively adjusts its output to more closely approximate the protected expression copied from the training dataset. The LLM records the results of this process in a large set of numbers called "weights" that are stored within the model. These weights are entirely and uniquely derived from the protected expression in the training dataset. Generally, the more data the LLM copies during training, the better the LLM's ability to simulate the protected expression within that data as part of the LLM's output. Moreover, generally speaking, the more diverse the training data is in prose, format, or subject matter, and the higher the quality of the written works that are within the training data, the better the LLM will be at producing high quality textual

output. As a result, high-quality textual works, including longform creative works, are highly valuable to companies seeking to compile AI training datasets.

30. Once the LLM has copied and ingested the textual works in the training dataset and converted the protected expression into stored weights, the LLM can emit convincing simulations of natural written language in response to user prompts. Whenever an LLM generates a response to a user prompt, it is performing a computation that relies on these stored weights, with the goal of imitating the protected expression ingested from the training dataset.

31. A significant portion of the material in the datasets that Apple downloaded and used for the purposes of training its LLMs consists of copyrighted works—including books written by Plaintiffs and Class Members—that Apple copied without consent and without providing credit or compensation.

B. The OpenELM language models were trained on copyrighted works.

32. In April 2024, Apple first announced the availability of the OpenELM language models on its website: "[W]e release OpenELM, a state-of-the-art open language model. OpenELM uses a layer-wise scaling strategy to efficiently allocate parameters within each layer of the transformer model, leading to enhanced accuracy."

33. The set of OpenELM language models released in April 2024 included variants called OpenELM-270M, OpenELM-450M, OpenELM-1_1B, and OpenELM-3B. The main difference between these models is the parameter size; a larger parameter size means the model can store more weights and perform more complex tasks (requiring more computing power). For instance, Apple's OpenELM-3B language model is so named because the model stores three billion ("3B") weights derived from protected expression found in its training dataset.

34. Each OpenELM model is hosted on a website called Hugging Face, where it has a "model card," a file accompanying an AI model that typically describes the model, its intended uses and limitations, its training parameters, and the training dataset used to train the model. The model card for each OpenELM model states: "Our pre-training dataset contains RefinedWeb, deduplicated PILE, a subset of RedPajama, and a subset of Dolma v1.6, totaling approximately 1.8 trillion tokens."

35. Apple's GitHub repository confirms that "OpenELM was pretrained on public datasets. Specifically, our pre-training dataset contains RefinedWeb, PILE, a subset of RedPajama, and a subset of Dolma v1.6."

36. Apple also published a paper about OpenELM ("OpenELM Paper"). In a table called "Dataset used for pre-training OpenELM," shown at right, Apple reveals that a large quantity of training data comes from the "Books" subset of a dataset called "RedPajama." The OpenELM Paper does not further describe the contents of the RedPajama dataset.

Source	Subset	Tokens
RefinedWeb		665 B
RedPajama	Github	59 B
	Books	26 B
	ArXiv	28 B
	Wikipedia	24 B
	StackExchange	20 B
	C4	175 B
PILE		207 B
Dolma	The Stack	411 B
	Reddit	89 B
	PeS2o	70 B
	Project Gutenberg	6 B
	Wikipedia + Wikibooks	4.3 B

Table 2. Dataset used for pre-training OpenELM.

37. But information about the RedPajama dataset is available elsewhere. The RedPajama dataset is hosted on Hugging Face. According to the documentation for the RedPajama dataset that was available there until around April 2024, its "Books" component is a copy of the "Books3 dataset" that is "downloaded from Huggingface [sic]" when a user runs the script that automatically assembles the RedPajama dataset. Therefore, anyone who used the "Books" subset of the RedPajama dataset for training an AI model used a copy of the Books3 dataset. The documentation for the RedPajama dataset does not further describe the contents of Books3.

38. A paper describing the RedPajama dataset (the "RedPajama Paper") further states that Apple's OpenELM models were trained on RedPajama-V1, which was described as "a publicly available, fully open, best-effort reproduction of the training data . . . used to train the first iteration of LLaMA family of models." The RedPajama Paper's recitation of the components of the LLaMA training dataset includes six of the same components (with the same token counts) that are listed in Apple's paper on OpenELM: GitHub, Books, ArXiv, Wikipedia, StackExchange, and C4. The RedPajama Paper later confirms that the "Books" component, with 26B tokens, included Books3.

39. A description of the contents of the Books3 datasets is available elsewhere. Books3 is a component of a separate AI training dataset called the Pile that was curated by a research organization

called EleutherAI. In December 2020, EleutherAI introduced this dataset in a paper called "The Pile: An 800GB Dataset of Diverse Text for Language Modeling" ("The Pile Paper"). This paper describes the contents of Books3:

> Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker … Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset … We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.

40. Similarly, Shawn Presser, who compiled the Books3 dataset, publicly confirmed in 2020 that that dataset is composed of "all of Bibliotik" and includes .txt files containing the full text of approximately 196,640 books.

41. Bibliotik is one of several notorious "shadow library" websites. Others include Library Genesis (aka LibGen, Z-Library, or B-ok), Sci-Hub, and Anna's Archive. The AI-training community has long been interested in these shadow libraries because they host and distribute vast quantities of unlicensed copyrighted material. For that reason, these shadow libraries violate the U.S. Copyright Act.

42. The 196,640 books in the Books3 dataset exist in .txt file format. A .txt file (pronounced a "text" file) is a simple file format that stores text data without any formatting, fonts, or images. Accordingly, the Books3 dataset consists of the text of the underlying 196,640 books.

43. By using the entire text of each book in Books3, Apple used copies of entire works to train its OpenELM models.

44. Plaintiffs' Infringed Works are among the works in the Books3 dataset.

45. Until October 2023, the Books3 dataset was available from Hugging Face. At that time, the Books3 dataset was removed with a message that it "is defunct and no longer accessible due to reported copyright infringement."

46. Presser himself has acknowledged that "we almost didn't release the data sets at all because of copyright concerns."

47. Before October 2023, anyone who used the "Books" subset of the RedPajama dataset for training necessarily made a copy of the Books3 dataset. Based on the information revealed in the OpenELM research paper and on its model card on Hugging Face, this includes Apple.

48. In sum, Apple has admitted training its OpenELM large language models on a copy of the "Books" subset of the RedPajama dataset, which in turn is a copy of the Books3 dataset. Therefore, Apple trained its OpenELM models on a copy of Books3, a known body of pirated books.

49. Because Plaintiffs' Infringed Works are part of Books3, Apple trained OpenELM on one or more copies of the Infringed Works and directly infringed Plaintiffs' copyrights along with the copyrights of the Class.

C. The Apple Intelligence Foundation Language Models were trained on copyrighted works.

50. In June 2024, Apple announced its commercial AI software platform, called Apple Intelligence. Apple Intelligence includes several AI models developed by Apple. Several of these models are called the *Apple Intelligence Foundation Language Models*.

51. The Foundation Language Models were described in a paper of the same name released by Apple on July 29, 2024 (the "FLM Paper"). According to that paper, the models are multimodal, meaning they can process, understand, and generate multiple types of data, including text and images. It follows that the Foundation Language Models were trained on image datasets as well as sets of text data.

52. The adjective *foundation* is commonly used to describe AI models that have broad capabilities to perform a wide variety of tasks. Consistent with this, Apple describes its Foundation Language Models as "highly capable in tasks like language understanding, instruction following, reasoning, writing, and tool use … These foundation models are at the heart of Apple Intelligence."

53. The FLM Paper emphasizes the special importance of a foundation model's capacity to write: "[w]riting is one of the most critical abilities for large language models to have, as it empowers various downstream use[s]."

54. In the FLM Paper, Apple identifies two separate foundation language models: *AFM-server* and *AFM-on-device*. The AFM-server model is a larger model that is intended for use

through an Apple-operated cloud service called Private Cloud Compute. The AFM-on-device model, by contrast, is intended to be small enough to be used directly on Apple devices (e.g., iPhones and laptops). According to the FLM Paper, the AFM-on-device model is "initialize[d] … from a pruned 6.4B model (trained from scratch **using the same recipe as AFM-server**.)" (emphasis added). This means that both the AFM-server and AFM-on-device models are trained on the same corpus of training data.

55. In the FLM Paper, Apple reveals three sources of training data: "data we have licensed from publishers, curated publicly available or open-sourced datasets, and publicly available information crawled by our web-crawler, Applebot."

56. In describing the first source of training data—"data we have licensed"—Apple says only that it "identif[ied] and license[d] a **limited amount** of high-quality data from publishers" (emphasis added). In addition to being comparatively "limited" in quantity, Apple does not use this licensed data during the main phase of training the Foundation Language Models—what Apple calls "core pre-training"—but during a subsequent phase called "continued pre-training."

57. As to its second source of training data—"publicly-available or open-sourced datasets"—Apple does not elaborate on the specific datasets used, saying only, "We evaluated and selected a number of high-quality publicly-available datasets with licenses that permit use for training language models." Apple then filtered the datasets to remove personally identifiable information, copyright management information, and other undesired text before including them in the pre-training mixture.

58. In the parlance of AI training datasets, Apple's phrase "publicly available" is one commonly used to falsely conjure up the idea of works made publicly available by the author. But the "public" nature of a dataset does not mean that the data collected in the dataset was obtained lawfully or that the party providing copies of the dataset has authority to extend a valid license to use the underlying copyrighted works. When companies describing AI training datasets use the phrase "publicly available" to describe those datasets, it has in practice meant only that the datasets were downloaded somewhere from the public internet, which contains a vast number of copyrighted works

by authors and owners who have not granted a license for reproduction. There is a name for this kind of downloading of copies: copyright infringement. There is also a name for the infringing copies: pirated works.

59. For instance, Apple described its training data for OpenELM, including data from the Pile, as "public datasets," even though the Pile contains an extraordinary number of pirated copies of copyrighted works. As another example, Meta Platforms also trained its LLaMA language models on Books3, which it described as a "publicly available dataset for training large language models" despite the fact that none of the authors or owners whose works appear in Books3 ever consented to having their works included. Books3 was "publicly available" only in the limited sense that at one time, it could be acquired by anyone with an internet connection.

60. Similarly, in the context of AI training datasets, Apple's phrase "open source" is commonly used to falsely conjure up the idea of works made available by the author under a permissive copyright license (e.g., a Creative Commons license). In practice, what it really means is copies of works made freely available by someone other than the author, without the author's permission. Again, these are just pirated works.

61. For instance, EleutherAI, the group that created The Pile—the dataset that included Books3—described it as a "diverse, open source language modelling data set" even though the copyrighted works in the Books3 portion were included without authors' and owners' consent. Only the copyright holder can offer their copyrighted work to the public under an open-source license. A third party cannot usurp that right.

62. Therefore, when Apple says that a major source of the training data for its Foundation Language Models is "publicly-available or open-sourced datasets," this should, on information and belief, be read to mean "certain pirated works or certain other pirated works." Because Books3 has been described by people in the AI industry as a "publicly available" or "open-sourced" dataset, and because Apple already had a copy of Books3 that it had used for training its OpenELM models, on information and belief, Apple's reference to "publicly-available or open-sourced datasets" includes Books3, and Apple therefore included Books3 in its private data library, which formed the basis for the

training dataset for its Foundation Language Models. Thus, it is also, on information and belief, the case that the "curated publicly available or open-sourced datasets" that Apple copied to create a private data library for the purposes of training its Foundation Language Models (and future models) contain copyrighted material, including Plaintiffs' and Class Members' copyrighted works.

63. Plaintiffs' Infringed Works are part of Books3. It follows that, on information and belief, Apple trained its Foundation Language Models on one or more copies of the Infringed Works, thereby directly infringing the copyrights of the Plaintiffs.

64. On information and belief, Apple has created a permanent digital library—and likely multiple such libraries—containing copies of all of its "curated," purportedly "publicly-available or open-sourced" datasets, like Books3, in expectation of future uses.

65. In obtaining a copy or copies of Books3, Apple made unauthorized reproductions of Plaintiffs' Infringed Works.

66. As to a third source of training data—web pages crawled by Applebot—Apple says, "we crawl publicly available information using our web crawler, Applebot … and respect the rights of web publishers to opt out of Applebot." While Apple officially acknowledged Applebot's crawling of webpages in mid-2015, its activity had been spotted in logs as early as November 2014. Around June 2024, Apple revealed that it was using Applebot-scraped data for training its AI models. In response to this disclosure, by August 2024, numerous major commercial web publishers had chosen to opt out of Applebot training.

67. But Apple's Foundation Language Models had necessarily been trained well before the release of the FLM Paper describing them in July 2024. For that reason, Apple's disclosure in June 2024 that it was using Applebot data to train language models came too late for any of these opt-outs to matter. Apple had already scraped the data and trained language models with it. On information and belief, Apple has retained copies of all Applebot data scraped before this wave of opt-outs, in expectation of training future models, as part of its AI training-data library.

68. In a November 2024 paper by George Wukoson and Joey Fortuna called "The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLMs," the

authors studied LLM training datasets made by algorithmically filtering scraped web pages. The authors concluded that such "datasets are disproportionately composed of high-quality content owned by commercial publishers of news and media websites." In turn, this material is often covered by registered copyrights. Thus, on information and belief, the parts of Apple's private shadow library and training datasets that come from filtered Applebot pages includes copyrighted works from commercial news and media websites.

69. The shadow libraries that host millions of unlicensed copyrighted books are also part of the "publicly available information" reachable by a web scraper like Applebot. Hence, on information and belief, part of Apple's training-data library is sourced from shadow libraries via Applebot. On information and belief, Apple obscures the libraries and datasets used to train its Foundation Language Models to conceal its use of copyrighted materials. On information and belief, Apple's decision not to disclose the contents of its internal data libraries, or the training datasets for its Foundation Language Models, stems in part from the fact that Apple was the subject of negative press for using a subset of data from The Pile, containing captions from thousands of YouTube videos, at least in training the OpenELM models and, on information and belief, also in training the Foundation Language Models.

70. In obtaining copies of Plaintiffs' Infringed Works from web scrapes via Applebot, Apple made additional unauthorized reproductions of Plaintiffs' Infringed Works.

71. Apple's sources of training data also included purchased print copies of books that were scanned into Apple's digital systems; Apple's existing troves of books and audiobooks, including within Apple Books; and third parties' data archives, which Apple contracted to access. Like the pirated and scraped datasets, at least some of the texts in each of these categories and sources, too, were copied after acquisition and used for training as well as other purposes, including the creation of permanent data libraries, the performance of research, and the transfer of copies to third parties.

72. As part of its development of Apple Foundation Models (AFMs), also known as Foundation Language Models (FLMs), Apple created "teacher" models that generated synthetic data, "classifier" models that filtered models' inputs and outputs, applications based on fine-tuned versions

of those models, and multiple generations of the models themselves, including the third generation released in June 2026. On information and belief, the training data for these models and applications includes data from Apple's private data libraries, which encompass Plaintiffs' Infringed Works through datasets including, but not limited to, Books3 and Library Genesis.

73. Most recently, in June 2026, Apple publicly released to developers the third generation of its language models, now called Apple Foundational Models or "AFM." This update includes four new language models, called "AFM 3 Core," "AFM 3 Core Advanced," "AFM 3 Cloud," and "AFM 3 Cloud Pro."

74. In a press release, Apple describes the training data for its AFM 3 models: "To train our foundation models, we use a mixture of data that includes publicly available information, data licensed or purchased from third parties, open-sourced data, data obtained through dedicated studies, and synthetic data." These categories are essentially equivalent to the categories of training data disclosed in the FLM paper (see ¶ 55) with two new ones: "data obtained through dedicated studies" and "synthetic data." On information and belief, "synthetic data" refers to data generated by other Apple language models, including models that were trained on datasets including copies of Plaintiffs' Infringed Works such as Books3 and Library Genesis.

75. Apple executives like Craig Federighi and Amar Subramanya have stated that the third-generation of Apple Foundation Models were the product of Apple's collaboration with Google and were trained using proprietary data and refined using outputs from Gemini frontier models. On information and belief, these statements that Apple's new models are "trained using proprietary data" means that they were trained using data, including Plaintiffs' Infringed Works, that Apple had stored in its private data library after training the earlier Foundation Language Models.

D. Apple's models for data classification were trained on copyrighted works.

76. In the FLM Paper, Apple says that Applebot pages are "processed by a pipeline which performs quality filtering … using heuristics and model-based classifiers." In this context, the term "model-based classifier" refers to a separate AI model that has been trained to algorithmically rate the quality of and sort scraped web pages. On information and belief, the models that underlie these

model-based classifiers were themselves trained on datasets that include unlicensed copyrighted works, including Plaintiffs' Infringed Works.

E. All of Apple's LLMs and model-based classifiers likely used multiple shadow libraries.

77. As copyright-infringement litigation has proceeded against other companies training LLMs, it has become clear that these companies downloaded, torrented, and/or used more pirated works and shadow libraries than was publicly known at the time of the complaint. Across these suits, the additional shadow libraries implicated include Anna's Archive, Pirate Library Mirror ("PiLiMi"), Z-Library (also known as B-ok), LibGen, and Sci-Hub. *See, e.g.*, Joint Letter Brief, *In re OpenAI ChatGPT Litig.*, No. 3:23-cv-03223-AMO (N.D. Cal. Jan 17, 2025) (Dkt. No. 254, at 1, 5) (acknowledging OpenAI's production of datasets "LibGen 1" and "LibGen 2"); First Consolidated Amended Complaint, *Nazemian, et al. v. NVIDIA*, No. 4:24-cv-01454-JST (N.D. Cal. Jan. 16, 2026) (Dkt. No. 235, at 12-13) (describing communications between NVIDIA and Anna's Archive, and alleging pirating from LibGen, Z-Library, Pirate Library Mirror, and Sci-Hub); Order Denying the Plaintiffs' Motion for Partial Summary Judgment and Granting Meta's Cross-Motion for Partial Summary Judgment, *Kadrey, et al. v. Meta*, No. 3:23-cv-03417-VC (N.D. Cal. June 25, 2025) (Dkt. No. 598, at 11) ("In early 2024, Meta also downloaded Anna's Archive, a compilation of shadow libraries including LibGen, Z-Library, and others."); Order on Fair Use, *Bartz, et al. v. Anthropic*, No. 3:24-cv-05417-WHA (N.D. Cal. June 23, 2025) (Dkt. No. 231, at 3) (noting that Anthropic downloaded millions of copies of books from LibGen and PiLiMi). According to the administrators of Anna's Archive, "virtually all major companies building LLMs contacted us to train on our data. . . We have given high-speed access to about 30 companies." The piracy is so widespread as to constitute a *de facto* industry practice.

78. Apple, like the companies listed above, is also a major company competing to build LLMs. As there was with the companies above, there is publicly available information indicating that Apple downloaded copyrighted works and used them in the multiple ways alleged herein. Thus, on information and belief, Apple, like the companies listed above, also very likely accessed one or more of these notorious shadow libraries to obtain large volumes of illegal copies of copyrighted books. Like

other major technology companies that did the same thing, Apple, on information and belief, did so while avoiding paying the customary fee for books and without permission from the copyright holder. Apple's internal documents reveal that Apple's Artificial Intelligence Machine Learning ("AIML") Team used Books3 to train many, and perhaps all, of its large language models. In December 2023, an Apple engineer noted that "essentially all models" that Apple had created as of that time were trained on Books3.

79. Other internal documents cite to "libgen," which is shorthand for the shadow library Library Genesis. Like Books3, Library Genesis is made up of unlicensed copyrighted material. Library Genesis currently contains more than 7.5 million books and 81 million research papers.

80. In June 2023, Apple employees suggested that while they could not retrieve contents from Library Genesis "directly," they could likely find a "direct dump" of the contents on another website like Huggingface, where Apple employees sourced a copy of Books3.

81. In July 2025, Apple employees stated: "[s]ince our current research text cont pretrain data contains libgen, we have to regenerate one without libgen"; and "out [sic] current continued pretrain mixture includes libgen." Other employees responded "Yeah, unfortunately, we need to remove that for now," and "[w]e can train on it internally but can not release any models including that."

82. By downloading a copy or copies of Library Genesis, Apple made unauthorized copies of Plaintiffs' Infringed Works. These unauthorized copies constituted copyright infringement regardless of what, if anything, Apple later did with the books.

F. Apple integrates the Foundation Language Models, including via Apple Intelligence, into Apple's commercial products.

83. Apple has integrated its Apple Intelligence products, including its Foundation Language Models, into a wide variety of commercially available products, including versions of the iPhone 15, 16, and 17; the iPad, iPad mini, and iPad Pro; the MacBook Air, MacBook Pro, iMac, Mac mini, and Mac Studio; and Apple Vision Pro.

84. In addition, Apple has offered Apple Intelligence, and the FLM in particular, to developers of Apple device software applications, or "apps," as a way to "create new intelligence features" using "the on-device large language model" at the core of Apple Intelligence. Apps are software applications, designed to run on devices such as smartphones, tablets, and other devices, that permit end-users to accomplish certain tasks. With many apps available on Apple devices, the end-users are typically consumers. According to Apple's public-facing site for app developers, the "Foundation Models framework provides" developers "access to Apple's on-device large language model that powers Apple Intelligence," and can be used by developers to create apps that permit consumers to "generat[e] creative content" and complete "a diverse range of text generation tasks." Apple's site specifically highlights the ability of Apple's FLM on-device model to "[c]ompose creative writing," even including, as an example, the following prompt an app user could enter: "Generate a short bedtime story about a fox." Additionally, Apple has released code and a machine-learning framework to assist development of iOS apps that integrate the OpenELM models.

85. Unsurprisingly, within several months of the Foundation Language Models becoming available to Apple developers for iOS 26, apps have appeared in the Apple App Store that allow users to generate output that could, in the short- or long-term, compete with copyrighted works. For example, Apple touted the app "Lil Artist" as one app relying on FLM, describing it as "combin[ing] the capabilities of the Foundation Models framework and the ImageCreator API to customize illustrated stories for children." Another app available through the Apple App Store, called "Locally," includes a "write professionally" button that encourages creative, long-form content generation. These apps, similar apps that currently exist, and similar apps that will inevitably be created in the near future using the FLM models or an OpenELM model, can be used to generate output that could compete with, dilute the market for, or simply be perceived and consumed by consumers as substitutes for Plaintiffs' and Class Members' copyrighted works. Apple earns revenue from the distribution and use of these apps via its App Store.

86. In September 2026, the AFM 3 models were released to the public as part of Apple's OS 27 update.

87. Apple has reportedly explored a paid tier for users of its Apple Intelligence products, with some early reports suggesting it may do so as early as 2027.

G. Apple scanned copies of Plaintiffs' and Class Members' books and then made numerous subsequent copies for training and reuse in a central repository.

88. In 2025, Apple decided to begin to scan books as a method to obtain more textual data to use for training its models. Apple then launched "Project Borders," its project focused on digitizing books to use in training its AFM models. In July 2025, Apple noted it had around [redacted] books to scan. Apple then procured vendors to obtain and scan books, which provided Apple with digitized copies of books.

89. Apple decided to expand its book-scanning project after determining it was cheaper to buy, scan, and digitize books than to license copies of eBooks from publishers.

90. By October 2025, Apple was scanning [redacted] books per day and had a goal of scanning [redacted] books by the end of March 2026. On information and belief, Apple made additional copies of the books that it scanned, some of which were used for training, some of which were used for, and stored in Apple's central repositories for future uses including, research and transfer to third parties.

H. Apple built and still maintains central repositories of Plaintiffs' and Class Members' works for future uses.

91. While Apple sources massive amounts of data for training its large language models, it does not delete such data after training its models. Instead, Apple retains these datasets in private data repositories that remain available for future research and reuse.

92. Apple's internal documents show that Apple has amassed central repositories of electronic versions of books and other printed material from a variety of sources, including (1) book scanning; (2) book/journal licensing; (3) Internet Archive downloads; and (4) open web PDF downloads. As one Apple engineer stated, "We are really building Apple's repository of all the printed information in the world."

93. Documents describing Apple's process for scanning books indicate that process, in which vendors send scanned PDF versions of books to Apple, Apple validates and processes the scanned book data, and that data "is then centrally stored."

94. Apple uses a variety of methods, vendors, and processes for storing its data in central repositories. Apple's repositories of electronic versions of books include, but are not limited to, data buckets (repositories hosted by various third-party services) on [redacted] Amazon S3, Google Cloud Storage. Apple also maintains data repositories for its foundational model data and synthetic data on Github.

95. Apple makes its central repositories accessible across teams, including those other than the AIML team responsible for designing and training Apple's AFM models. Internal documents show that Apple planned for all AFM datasets, models, and lineage information (i.e., where the data came from) to be "easily accessible" by various Apple teams, including AIML, Apple Services Engineering ("ASE"), Video Engineering ("VE"), and Software Engineering ("SWE").

96. Even when Apple identifies a dataset of textual works as being problematic for training publicly released large language models, Apple still maintains copies of those datasets in its central repositories for future research and additional uses.

97. On information and belief, Apple maintains copies of Plaintiffs' Infringed Works and Class Members' works in its central repositories for future uses.

I. Apple's conduct impairs the markets for Plaintiffs' and Class Members' works.

98. Apple has neither paid nor sought permission from Plaintiffs for the use of their copyrighted works. Instead of doing so, Apple downloaded, torrented, scraped, or otherwise copied vast quantities of copyrighted works—including illegally compiled datasets such as Books3—that included Plaintiffs' works like the Infringed Works. In so doing, Apple has, first, deprived Plaintiffs of the revenue that would have been generated had Apple approached Plaintiffs or their licensing agents directly to license copies of their works. Plaintiffs and similarly situated creators previously licensed their work for their own commercial uses.

99. There are numerous examples of publicly reported AI licensing deals. For example, Microsoft reportedly reached an agreement with HarperCollins to license a select number of non-fiction titles to train AI models for just three years; compensation for participating authors, who have to opt-in before their works can be licensed, is $2,500 per work. Wiley likewise reported in 2025 that it had executed multiple AI licensing agreements totaling $40 million. Myriad licensing systems have been launched and are continuing to develop, including the Copyright Clearance Center's collective AI licensing scheme and the Created by Humans licensing platform. Further, several AI dataset licensing companies have formed a trade group called the Dataset Providers Alliance in order to promote standardization and transparency in the licensing of intellectual property content for AI and machine learning datasets. Currently, some researchers estimate, the AI training license market is valued at approximately $2.5 billion; within a decade, it may close in on nearly $30 billion.

100. Apple itself understands the value of copyrighted works and the market that exists for paying creators to use their works for training. For instance, it struck an agreement with Shutterstock to "use hundreds of millions of images, videos and music files" valued between an estimated $25 to $50 million. Similarly, Apple has contacted news organizations like Condé Nast, NBC News, and IAC to license news article archives. Nonetheless, Apple has not compensated Plaintiffs' and Class Members whose works it copied and used in training its models.

101. Apple has not only directly deprived Plaintiffs and Class Members of the licensing revenues Plaintiffs and Class Members could have earned from Apple for Apple's use of their works to train their LLMs. Apple has also contributed, through its conduct, to the impairment of the emergence of lawful licensing regimes. Such nascent licensing regimes are more successful if large corporate licensees participate early and at scale, contributing significant licensing revenues and stabilizing the demand side of such a licensing market. Apple's choice to instead copy pirated works on a massive scale has, merely through the absence of a corporate cornerstone of such a market, blunted the development of an AI training dataset licensing regime.

102. Apple's conduct has also depressed the demand for licensing from other potential licensees by normalizing and encouraging other potential licensees not to license works for AI training.

Apple has significant consumer brand loyalty and is one of the most valuable companies in the world. Smaller potential licensees are less likely to pursue licensing if they see that Apple, with its cadre of retained and in-house lawyers, has determined it does not need to do so. Smaller potential licensees are also less likely to pursue licensing if they believe they cannot do so and remain competitive with larger companies with more capital (like Apple) who have declined to pay licensing fees and instead flouted copyright law. Finally, smaller potential licensees are less likely to pursue licensing if Apple's conduct has normalized the use of pirated works, which use Apple has largely attempted to hide from consumers.

103. Second, Apple's unauthorized use of Plaintiffs' copyrighted works to train the OpenELM models and Foundation Language Models has also caused and threatens to cause substantial harm to the actual and potential markets for those works. As described above, Apple has developed, designed, created, maintained, and/or distributed functionalities and applications that are designed to, and can, generate works that imitate or otherwise resemble the Infringed Works. Apple has also encouraged and facilitated third parties' design, creation, maintenance, and/or distribution of relevantly similar applications. All of these unauthorized uses of the Infringed Works creates a risk of market dilution, which may "lead to a loss of sales" by "harm[ing] the market for access to those works." Works generated using the OpenELM models or the Foundation Language Models will inevitably start competing with Plaintiffs' works (like the Infringed Works) and ultimately dilute royalty pools as AI-generated output increasingly floods the market. Already, "low-quality sham 'books'" created by other LLMs have begun overwhelming the market as scammers generate "unauthorized 'biographies' of authors that are simply AI-generated rehashings of their lives, often based on autobiographical works." Other scams include "companion books" that summarize the key points from the original novel, with "little to no original analysis or commentary and are meant only to confuse consumers and skim sales off of the real books." These works have already entered book marketplaces like Amazon.

104. Apple's models—whether integrated in Apple Intelligence, an Apple-developed app, or a third-party-developed app—are designed to, and can, generate outputs that may substitute the kinds of expressive written work that Plaintiffs are hired to produce, potentially diminishing demand for

books and human-produced stories. Plaintiffs face potential ongoing harms through lost publication opportunities, reduced recognition, and lost sales, among other harms.

105. This type of substitution of human-created works for AI-generated works has already been documented in creative markets. For example, the *New York Times* reported in a recent front-page story: "Publishers and authors worry that books by real writers are getting lost in the sea of digital slop, as A.I.-enabled novels flood the market." According to one best-selling novelist quoted in that article: "It bogs down the publishing ecosystem that we all rely on to make a living . . . It makes it difficult for newer authors to be discovered, because the swamp is teaming [sic] with crap." Although some stigma currently attaches to known AI-generated novels, another writer predicts, "[e]ventually . . . readers will not care." As an additional example, researchers have recently shown that some readers prefer the outputs of AI models trained on copyrighted works over the work of human writers.

106. Additionally, in a 2025 working paper analyzing the impact of generative AI on the stock image market, the authors describe their findings as follows:

> In this paper, we ask three questions with the aims of shedding light on how GenAI will change markets for creative goods and informing the active debate on copyright and fair use.
>
> **First, to what extent do GenAI goods compete with—or become substitutes for—non-GenAI goods? We provide causal evidence that consumers view GenAI-produced goods as substitutes for traditional creative goods. Thus GenAI goods are competitive with non-GenAI goods.** Further, the market expands as GenAI goods enter the market, with GenAI goods consumption more than replacing non-GenAI goods.
>
> **Next, does GenAI result in crowd-out of non-GenAI firms and goods?** Policymakers in particular have expressed this concern given the substantial differences in production costs brought by GenAI. **Our results show that GenAI does crowd out non-GenAI firms and goods. Across almost all markets, we see substantial entry of GenAI firms and exit of non-GenAI firms, with net production expansion, but a decrease in any given product's probability of sale. We document that this concern is particularly relevant for niche markets, where we find production tipping almost completely to GenAI.** . . .
>
> . . .
>
> The substitution to non-GenAI content, and significant exit of non-GenAI artists, provide evidence of crowd-out of non-GenAI content. **In the long run, this may lead to a decline in the production of novel original content for model training, and even to a decline in non-GenAI content altogether. Further, we document a reduction in sales rates for artists.** While these reductions may be partially compensated for by the increase in production, not all artists, and in particular not all

non-GenAI artists, will be able to fully compensate via scale, and thus their long-run profitability may be at risk. **Thus, our results provide some validity to concerns that GenAI content will both replace non-GenAI content and greatly diminish the value of the original content, affecting the fourth pillar of fair use.**

VI. CLASS ACTION ALLEGATIONS

107. The "Class Period" as defined in this Complaint begins at least three years before September 5, 2025 and runs through the present.

108. As used here, the term "Apple Uses" refers collectively to Apple's numerous separate infringing uses of Plaintiffs' and Class Members' works, including the Infringed Works, as set forth below:

A. *Acquiring Data via Pirated Sources:* Apple's unauthorized reproduction of Plaintiffs' and Class Members' works, including the Infringed Works, as part of Apple's downloading, torrenting, or other mode of initial acquisition of one or more datasets of pirated works, including Books3, Library Genesis, Z-Library, Anna's Archive, Sci-Hub, Internet Archive, and/or Pirate Library Mirror;

B. *Acquiring Data via Applebot:* Apple's unauthorized copying of Plaintiffs' and Class Members' works, including the Infringed Works, as part of Apple's scraping of the public internet using Applebot or other methods;

C. *Reproducing Data after Acquisition:* Apple's unauthorized copying of Plaintiffs' and Class Members' works, including the Infringed Works, whether for the purpose of training, research, transfer, or permanent retention, regardless of the mode of acquisition (including but not limited to downloading, scraping, licensing, contracting, and scanning);*Curating:* Apple's unauthorized reproduction and use of Plaintiffs' and Class Members' works, including the Infringed Works, when filtering, cleaning, and processing at least one pirated dataset, such as Books3, and components of such datasets, to create a curated, private shadow library of datasets for use in various stages of training LLMs;

D. *Training OpenELM Models:* Apple's unauthorized reproduction and use of Plaintiffs' and Class Members' works, including the Infringed Works, when training its OpenELM models;

E. *Training FLMs:* Apple's unauthorized reproduction and use of Plaintiffs' and Class Members' works, including the Infringed Works, when training all versions and generations of its Foundation Language Models;

F. *Training Classifier and Teacher Models:* Apple's unauthorized reproduction and use of unlicensed copyrighted works when training classifier and teacher models; and

109. *Retaining for Private Data Library:* Apple's unauthorized retention, curation, reproduction, and use of Plaintiffs' and Class Members' works, including the Infringed Works, along with all the data Apple has illegally downloaded, curated, reproduced, and used thus far, in the form of a private data library for use in models currently under development as well as future models—an AI data library that includes the Books3 dataset, which, in turn, includes the Infringed Works. **Class definition.** Plaintiffs bring this action for damages and injunctive relief as a class action under Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3), on behalf of the following Class:

> **All beneficial or legal owners of a registered United States copyright in any work that was 1) used in one or more of the Apple Uses during the Class Period, and 2) was registered with the United States Copyright Office a) within five years of the work's publication and before its use in one or more of the Apple Uses, or b) within three months of publication.**

110. This Class definition excludes: (a) Defendant; (b) Any of Defendant's parent companies, subsidiaries, and affiliates; (c) Any of Defendant's officers, directors, management, employees, subsidiaries, affiliates, or agents; (d) All governmental entities; and (e) The judges and chambers staff in this case, as well as any members of their immediate families.

111. **Numerosity: Federal Rule of Civil Procedure 23(a)(1).** The Class Members are so numerous and geographically dispersed that individual joinder of all Class Members is impracticable. The exact number of Class Members is currently unknown to Plaintiffs, as this information is in Defendant's exclusive control. On information and belief, there are more than several thousand

members in the Class across the United States. Accordingly, joinder of all Class Members in prosecuting this action is impracticable.

112. The Class can be identified, in part, through tools that allow a user to search for web domains included in the RedPajama dataset, the Pile dataset, and other datasets used for one or more of the Apple Uses.

113. The Class can further be identified by obtaining business records maintained by Apple, including the content of its private pirated libraries and datasets used for both its OpenELM models and Foundation Language Models.

114. **Typicality: Federal Rule of Civil Procedure 23(a)(3).** Plaintiffs' claims are typical of the claims of Class Members because Plaintiffs and all members of the Class were damaged by the same course of conduct of Defendant. Further, the relief sought is common to all Class Members.

115. **Adequacy of Representation: Federal Rule of Civil Procedure 23(a)(4).** Plaintiffs will fairly and adequately represent the interests of the members of the Class because Plaintiffs have experienced the same harms as the members of the Class and have no conflicts with any other members of the Class. Further, Plaintiffs have retained competent counsel who are experienced in litigating federal class actions and other complex litigation involving sophisticated, state-of-the art technology.

116. **Commonality and Predominance: Federal Rules of Civil Procedure 23(a)(2) and 23(b)(3).** Numerous questions of law and fact are common to each Class Member arising from Defendant's conduct, including:

A. Whether Apple acquired or downloaded reproductions of datasets, including but not limited to the RedPajama and Books3 datasets, that included Plaintiffs' and Class Members' copyrighted works;

B. Whether Apple copied Plaintiffs' and Class Members' works as part of Apple's scraping of the public internet using Applebot or other methods;

C. Whether Apple included Plaintiffs' and Class Members' works in any of the datasets used by Apple to train its OpenELM models and Foundation Language Models;

D. Whether Apple's inclusion of Plaintiffs' and Class Members' works in those datasets constituted or required the works' reproduction by Apple;

E. Whether Apple lacked authorization to reproduce copies of Plaintiffs' and Class Members' works;

F. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and used them in training its OpenELM models;

G. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and used them in training its Foundation Language Models;

H. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and retained them in a curated, private shadow library for Apple's use in training other models;

I. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and reproduced them for purposes of research;

J. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and reproduced and transferred them to third parties;

K. Whether this Court should enjoin Defendant from engaging in the unlawful conduct alleged herein;

L. Whether any affirmative defense excuses Defendant's conduct, including the fair use doctrine; and

M. Whether Apple's infringement was willful.

117. These and other questions of law and fact are common to the Class and predominate over questions affecting Class Members on an individual basis.

118. **Predominance & Superiority: Federal Rule of Civil Procedure 23(b)(3).** Defendant has acted on grounds generally applicable to the Class. A class action is superior to alternatives for the fair and efficient resolution of this controversy. Allowing the claims to proceed on a class basis will eliminate the possibility of repetitive litigation. Further, injunctive relief is appropriate with respect to the entire Class.

119. **Risk of Prosecuting Separate Actions.** The alternative of separate actions by individual Class Members risks inconsistent adjudications and is an inefficient use of limited judicial resources.

VII. CLAIM

DIRECT COPYRIGHT INFRINGEMENT — 17 U.S.C. § 501

120. Plaintiffs incorporate by reference all other allegations in this complaint.

121. As the owners of the registered copyrights in the Infringed Works and other copyrighted works, Plaintiffs and Class Members hold the exclusive rights to those works under 17 U.S.C. § 106. Plaintiffs and the Class Members never authorized Apple to make copies of their Infringed Works and other copyrighted works, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works), or exploit any other right exclusively reserved to Plaintiffs and the Class Members under the U.S. Copyright Act.

122. In performing the Apple Uses, including by downloading the works and making additional copies during various stages of training and development of the OpenELM models and the Foundation Language Models, Apple made all its copies of the Infringed Works and other copyrighted works without Plaintiffs' or Class Members' permission, violating their exclusive rights under the U.S. Copyright Act. Indeed, "the person who copies the textbook from a pirate site has infringed already, full stop." *Bartz et al. v. Anthropic*, 787 F. Supp. 3d 1007, 1025 (N.D. Cal. 2025). Regardless of how Apple uses the works in its private training-data library in the future, this cannot negate that the initial copying of works sourced from shadow libraries infringed on Plaintiffs' and Class Members' exclusive rights.

123. Plaintiffs and Class Members have been injured by Apple's acts of direct copyright infringement of the Infringed Works. Plaintiffs and Class Members are entitled to statutory damages, actual damages, restitution of profits, destruction of the infringing copies and models, and other remedies provided by law.

124. Apple's violation of Plaintiffs' and Class Members' exclusive rights was willful. Apple knew the datasets it downloaded, copied, stored, and trained its LLMs on contained copyrighted works.

VIII. PRAYER FOR RELIEF

125. Plaintiffs demand judgment on their behalf and on behalf of the Class against each Defendant as follows:

A. Allowing this action to proceed as a class action, with Plaintiffs serving as Class Representatives, and with Plaintiffs' counsel as Class Counsel;

B. Awarding Plaintiffs and the Class statutory damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity;

C. Permanently enjoining Defendant from the unlawful, unfair, and infringing conduct alleged herein;

D. Ordering destruction under 17 U.S.C. § 503(b) of all Apple LLMs that have ingested at least one copy of any of Plaintiffs' and Class Members' works;

E. Ordering destruction under 17 U.S.C. § 503(b) of all copies of those works that are maintained in Apple's private libraries and datasets;

F. An award of costs, expenses, and attorneys' fees as permitted by law; and

G. Such other or further relief as the Court may deem appropriate, just, and equitable.

IX. DEMAND FOR JURY TRIAL

Plaintiffs demand a jury trial for all claims.

DATED this 2nd day of November, 2026.

KELLER ROHRBACK L.L.P.

By *s/ William K. Dreher*
William K. Dreher, *Pro Hac Vice*
Benjamin Gould (SBN 250630)
Derek W. Loeser, *Pro Hac Vice*
Chris N. Ryder, *Pro Hac Vice*
1201 Third Avenue, Suite 3400
Seattle, WA 98101-3268
(206) 623-1900
Fax (206) 623-3384
bgould@kellerrohrback.com
wdreher@kellerrohrback.com
dloeser@kellerrohrback.com
cryder@kellerrohrback.com

SUSMAN GODFREY L.L.P.

By *s/ Rohit D. Nath*
Rohit D. Nath (SBN 316062)
Justin A. Nelson, *Pro Hac Vice*
Alejandra C. Salinas, *Pro Hac Vice*
Jordan Connors, *Pro Hac Vice*
Eleanor Runde, *Pro Hac Vice*
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
(310) 789-3100
rnath@susmangodfrey.com
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com
jconnors@susmangodfrey.com
erunde@susmangodfrey.com

BUTTERICK LAW PC

By *s/ Matthew Butterick*
Matthew Butterick (SBN 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
mb@buttericklaw.com

Attorneys for Plaintiffs

Exhibit A

Exhibit A

Registration Number	Title of Work	Author of Work
TX0008563355	A Love Like This	Eboni Manning aka Kianna Alexander
TX0009297307	After Hours Agenda	Eboni Manning aka Kianna Alexander
TX0009100320	After Hours Attraction	Eboni Manning aka Kianna Alexander
TX0009021729	After Hours Redemption	Eboni Manning aka Kianna Alexander
TX0009227221	After Hours Temptation	Eboni Manning aka Kianna Alexander
TX0009170932	Back to Your Love	Eboni Manning aka Kianna Alexander
TX0009341367	Can’t Let Her Go	Eboni Manning aka Kianna Alexander
TX0009154431	Can’t Resist Her	Eboni Manning aka Kianna Alexander
TX0009135026	Carolina Built	Eboni Manning aka Kianna Alexander
TX0008668191	Couldn’t Ask for More	Eboni Manning aka Kianna Alexander
TX0008217057	Every Beat Of My Heart	Eboni Manning aka Kianna Alexander
TX0008744613	Forever With You	Eboni Manning aka Kianna Alexander
TX0008672549	Love for All Time	Eboni Manning aka Kianna Alexander
TX0008979954	Never Let Me Go	Eboni Manning aka Kianna Alexander
TX0008448578	Tempo Of Love	Eboni Manning aka Kianna Alexander
TX0008787155	Then Came You	Eboni Manning aka Kianna Alexander

Exhibit A

Registration Number	Title of Work	Author of Work
TX0008193076	This Tender Melody	Eboni Manning aka Kianna Alexander
TX0009242549	What Happens After Hours	Eboni Manning aka Kianna Alexander
TX0009327754	Ankle Snatcher	Grady Hendrix
TX0009281005	BadAsstronauts	Grady Hendrix
TX0007957438	Horrorstor: A Novel	Grady Hendrix
TX0009231440	How to Sell a Haunted House	Grady Hendrix
TX0008276604	My Best Friend's Exorcism	Grady Hendrix
TX0008461565	Paperbacks from Hell: The Twisted History of '70s and '80s Horror Fiction	Grady Hendrix
PAu002397892	Satan Loves You	Grady Hendrix
TX0009212410	The Final Girl Support Group	Grady Hendrix
TX0008888592 and TX0008986901	The Southern Book Club's Guide to Slaying Vampires: A Novel	Grady Hendrix
TX0008645317 and TX0008645012	We Sold Our Souls: a Novel	Grady Hendrix
TX0009505624	Witchcraft for Wayward Girls	Grady Hendrix
TX0008809460	My Heart Struck Sorrow	John Hornor Jacobs
TX0008114163	The Conformity	John Hornor Jacobs
TX0008830088	The Sea Dreams It Is The Sky	John Hornor Jacobs
TX0007905434	The Shibboleth	John Hornor Jacobs
TX0007561564	This Dark Earth	John Horner Jacobs
TX0002595154	A Pride of Princes	Jennifer Roberson
TX0003445261	A Tapestry of Lions	Jennifer Roberson
TX0005388531	Children of the Lion	Jennifer Roberson

Exhibit A

Registration Number	Title of Work	Author of Work
TX0002655076	Daughter of the Lion	Jennifer Roberson
TX0006608590	Deepwood: Karavans 2	Jennifer Roberson
TX0003038048	Flight of the Raven	Jennifer Roberson
TX0004575188	Highwaymen: Robbers & Rogues	Jennifer Roberson
TX0006494921	Karavans	Jennifer Roberson
TX0005700486	Lady of Sherwood	Jennifer Roberson
TX0003467465	Lady of the Forest (A Novel of Sherwood)	Jennifer Roberson
TX0004292057	Lady of the Glen	Jennifer Roberson
TX0002124929	Legacy of the Sword	Jennifer Roberson
TX0005388529	Legacy of the Wolf	Jennifer Roberson
TX0008936965	Life and Limb	Jennifer Roberson
TX0004840105	Mad Jack	Jennifer Roberson
TX0005450939 and TX0005450940	Out of Avalon	Jennifer Roberson
TX0006141041	Piece of Mind	Jennifer Roberson
TX0004248401	Return to Avalon: a celebration of Marion Zimmer Bradley	Jennifer Roberson
TX0002040293	Royal Captive	Jennifer Roberson
TX0001405603	Shapechangers	Jennifer Roberson
TX0005348007	Shapechanger’s Song	Jennifer Roberson
TX0009100077	Sinners and Saints	Jennifer Roberson
TX0001570167	Smoketree	Jennifer Roberson
TX0009259320	Sword-Bearer	Jennifer Roberson
TX0004752101	Sword-Born	Jennifer Roberson
TX0007724532	Sword-Bound	Jennifer Roberson

Exhibit A

Registration Number	Title of Work	Author of Work
TX0003187269	Sword-Breaker	Jennifer Roberson
TX0002125935	Sword-Dancer	Jennifer Roberson
TX0003031154	Sword-Maker	Jennifer Roberson
TX0002585643	Sword-Singer	Jennifer Roberson
TX0005504873	Sword-Sworn	Jennifer Roberson
TX0002496298	The Court of the Summer King	Jennifer Roberson
TX0005556700	The Lion Throne	Jennifer Roberson
TX0001840915	The Song of Homana	Jennifer Roberson
TX0007614823	The Wild Road	Jennifer Roberson
TX0002138857	Track of the White Wolf	Jennifer Roberson

William Dreher (*pro hac vice*)
Derek W. Loeser (*pro hac vice*)
Cari C. Laufenberg (*pro hac vice*)
Benjamin Gould (SBN 250630)
KELLER ROHRBACK L.L.P.
1201 Third Avenue, Suite 3400
Seattle, WA 98101-3268
(206) 623-1900
wdreher@kellerrohrback.com
dloeser@kellerrohrback.com
claufenberg@kellerrohrback.com
bgould@kellerrohrback.com

Matthew Butterick (SBN 250953)
BUTTERICK LAW PC
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
(323) 968-2632
mb@buttericklaw.com

Rohit D. Nath (SBN 316062)
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
(310) 789-3100
RNath@susmangodfrey.com

Justin A. Nelson (*pro hac vice*)
Alejandra C. Salinas (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
(713) 653-7802
JNelson@susmangodfrey.com
ASalinas@susmangodfrey.com

Additional Counsel Listed in Signature Block

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
OAKLAND DIVISION

GRADY HENDRIX ET AL.,

Individual and Representative Plaintiffs,

v.

APPLE INC.,

Defendant.

This Document Relates To:
All Consolidated Actions

No. 4:25-cv-07558-YGR

FIRST AMENDED CONSOLIDATED CLASS ACTION COMPLAINT

FILED UNDER SEAL

JURY TRIAL DEMANDED

Plaintiffs Grady Hendrix, Jennifer Roberson, John Hornor Jacobs, and Eboni ~~McKinnon~~ Mckinnon Smith (a/k/a Kianna Alexander) (together "Plaintiffs"), on behalf of themselves and all others similarly situated (the "Class," as defined below), bring this class-action complaint ("Complaint") against Defendant Apple Inc. ("Apple" or "Defendant").

I. INTRODUCTION

1. Apple, one of the world's most valuable companies, has invested substantial capital and engineering resources into developing generative artificial intelligence ("AI") programs and technologies. Among Apple's AI products is Apple Intelligence. Apple regards Apple Intelligence as a breakthrough innovation that will make its users' experiences "profoundly different" across various product applications. Through Apple Intelligence, Apple hopes to add trillions to its market capitalization in coming years.

2. But Apple is building this new enterprise in part by using Books3, a dataset of pirated copyrighted books that includes the published works of Plaintiffs and the Class. Apple has developed, designed, maintained, and commercialized multiple large language models ("LLMs"). An LLM is a type of machine-learning model designed to emit convincingly naturalistic text outputs in response to user prompts. Companies developing LLMs "train" the LLMs to produce these outputs by feeding the models copies of an enormous quantity of data—frequently including massive numbers of textual works. Apple used the dataset called Books3 to train one set of LLMs, the OpenELM language models. Apple also, on information and belief, trained a second set of LLMs, its Foundation Language Models, using this same pirated dataset.

3. Apple's generative AI also uses Applebot, a software program that copies mass quantities of webpages (also known as "scraping"). Apple scraped data with Applebot for nearly nine years before disclosing that it intended to train its commercial AI products on this scraped data. Scrapers like Applebot can also reach so-called "shadow libraries" that host millions of other unlicensed copyrighted books, including, on information and belief, Plaintiffs' and Class Members' copyrighted works. Shadow libraries are the approximate equivalent of Napster for books; users upload unauthorized copies of books, and the shadow library helps other users download and access the works for free without paying the customary price. As has become public in other cases, a number of Apple's competitors in the AI race—including OpenAI, Meta Platforms, Anthropic, and NVIDIA—resorted to obtaining copyrighted materials from shadow libraries like Library Genesis, Z-Library, Anna's Archive, Sci-Hub, and Pirate Library Mirror. Upon information and belief, like its competitors, Apple

obtained works from these notorious pirate libraries—accessed either through Applebot or torrenting—to build its own permanent library of stolen material for training its models, for commercial research, and for other purposes.

~~3.~~4. Apple also illicitly obtained copyrighted books for AI training by other means, including by scanning books. After acquiring those books, Apple made additional copies of them. In some cases, Apple converted the books to new formats *and* made additional copies—for instance, when converting text to audio, and vice versa,~~;~~ and when generating synthetic data (that is, —data created by a generative AI model)—from models trained on books. And in all of these instances, Apple stored such data in central repositories and made such data available for additional future uses, including to make training more convenient.

~~4.~~5. The quality of Apple's models depends on the data on which it is trained. LLMs operate by copying and later simulating creative expression found in copyrighted works. For this reason, the inclusion of expressive, high-quality material—especially copyrighted material—in Apple's AI training datasets is deliberate and commercially significant. For instance, to access even more copyrighted material to develop its valuable generative AI products, Apple entered into a multimillion-dollar licensing agreement with Shutterstock. But not with Plaintiffs or the Class.

~~5.~~6. Plaintiffs and the Class are owners of registered copyrights for their published works. They did not consent to the use of their works in any model developed by Apple, including the Foundation Language Models and OpenELM language models.

~~6.~~7. The licensing market for AI training data is robust and growing. For instance, in February 2026, Microsoft launched a data-licensing marketplace called the Publisher Content Marketplace; Amazon has held discussions with publishers about building a similar service. Nevertheless, Apple did not compensate creators for use of their copyrighted works and concealed the sources of the datasets they downloaded, curated, and used for AI training to evade legal scrutiny. On information and belief, Apple continues to retain its own curated, private shadow library of pirated books to train AI models currently or soon to be under development and for other commercial purposes, all without seeking Plaintiffs' or Class Members' consent or providing them compensation.

8. Apple has copied the copyrighted works of Plaintiffs and the Class to train AI models whose outputs compete with and dilute the market for those very works. Without those works, Apple's generative AI models, Apple products that integrate those models via Apple Intelligence, and Apple ecosystems in which third-party products integrate those models (such as the Apple App Store) would have far less commercial value. This conduct has deprived Plaintiffs and the Class of control over their work, undermined the economic value of their labor, and positioned Apple to achieve massive commercial success through unlawful means.

9. In sum, Apple engaged in at least the following eight infringing uses as to Plaintiffs and the Class, as detailed in the remainder of this Complaint:

A. *Acquiring Data via Pirated Sources:* Apple made unauthorized reproductions of Plaintiffs' works as part of Apple's downloading, torrenting, or other mode of initial acquisition of one or more datasets of pirated works, including at least Books3 and Library Genesis.

B. *Acquiring Data via Applebot:* Apple made unauthorized copies of Plaintiffs' works as part of Apple's scraping of the public internet using Applebot or other methods;

C. *Reproducing Data after Acquisition:* Apple made unauthorized copies of Plaintiffs' works whether for the purpose of training, research, transfer, or permanent retention, regardless of the mode of acquisition (including but not limited to downloading, scraping, licensing, contracting, and scanning);

D. *Curating:* Apple made unauthorized copies of Plaintiffs' works when filtering, cleaning, and processing at least one pirated dataset, such as Books3, and components of such datasets, to create a curated, private shadow library of datasets for use in various stages of training LLMs;

E. *Training OpenELM Models:* Apple made unauthorized copies of Plaintiffs' works when training its OpenELM models;

F. *Training FLMs:* Apple made unauthorized copies of Plaintiffs' works when training all versions and generations of its Foundation Language Models;

G. *Training Classifier and Teacher Models:* Apple made unauthorized copies of Plaintiffs' works when training classifier and teacher models; and

~~A.~~H. *Retaining for Private Data Library:* Apple engaged in unauthorized retention, curation, reproduction, and use of Plaintiffs' works along with all the data Apple has illegally downloaded, curated, reproduced, and used thus far, in the form of a private data library for use in models currently under development as well as future models—an AI data library that includes the Books3 and Library Genesis dataset, which, in turn, includes Plaintiffs' works.

II. JURISDICTION AND VENUE

~~7.~~10. This Court has subject-matter jurisdiction under 28 U.S.C. § 1331 because this case arises under the Copyright Act (17 U.S.C. § 501).

~~8.~~11. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Apple is headquartered in this district.

III. DIVISIONAL ASSIGNMENT

~~9.~~12. Under Civil Local Rule 3-2(c), assignment of this case to the Oakland Division is proper because this case pertains to intellectual-property rights, which is a district-wide case category under General Order No. 44, and therefore venue is proper in any courthouse in this District.

IV. PARTIES

A. Plaintiffs

~~10.~~13. Plaintiff Grady Hendrix is an author who lives in New York. He owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "Hendrix Asserted Works"): *Horrorstor* (U.S. Copyright Office Registration Number TX0007957438); *My Best Friend's Exorcism* (USCO Registration Number TX0008276604); *Paperbacks from Hell* (USCO Registration Number TX0008461565); ~~and~~ *We Sold Our Souls* (USCO Registration Number TX0008645012); *Ankle Snatcher* (USCO Registration Number

TX0009327754); *BadAsstronauts* (USCO Registration Number TX0009281005); *How to Sell a Haunted House* (USCO Registration Number TX0009231440); *The Final Girl Support Group* (USCO Registration Number TX0009212410); *The Southern Book Club's Guide to Slaying Vampires: A Novel* (USCO Registration Number TX0008888592, TX0008986901); and *Witchcraft for Wayward Girls* (USCO Registration Number TX0009505624).

~~11.~~14. Plaintiff Jennifer Roberson is an author who lives in Arizona. She owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "Roberson Asserted Works"): *Life and Limb* (USCO Registration Number TX0008936965); *Shapechangers* (USCO Registration Number TX0001405603); *Legacy of the Sword* (USCO Registration Number TX0002124929); *Track of the White Wolf* (USCO Registration Number TX0002138857); *A Pride of Princes* (USCO Registration Number TX0002595154); *Daughter of the Lion* (USCO Registration Number TX0002655076); *Flight of the Raven* (USCO Registration Number TX0003038048); *A Tapestry of Lions* (USCO Registration Number TX0003445261); *Karavans* (USCO Registration Number TX0006494921); *Deepwood* (USCO Registration Number TX0006608590); *The Wild Road* (USCO Registration Number TX0007614823); *Lady of the Glen* (USCO Registration Number TX0004292057); *Lady of the Forest* (USCO Registration Number TX0003467465); *Lady of Sherwood* (USCO Registration Number TX0005700486); *Sword-Dancer* (USCO Registration Number TX0002125935); *Sword-Singer* (USCO Registration Number TX0002585643); *Sword-Maker* (USCO Registration Number TX0003031154); *Sword-Breaker* (USCO Registration Number TX0003187269); *Sword-Born* (USCO Registration Number TX0004752101); *Sword-Sworn* (USCO Registration Number TX0005504873); *Sword-Bound* (USCO Registration Number TX0007724532); *The Song of Homana* (USCO Registration Number TX0001840915); *Mad Jack* (USCO Registration Number TX0004840105); *Out of Avalon* (USCO Registration Number TX0005450939 & TX0005450940); *Sinners and Saints* (USCO Registration Number TX0009100077); *Smoketree* (USCO Registration Number TX0001570167); and *Sword-Bearer* (USCO Registration Number TX0009259320).

~~12.~~15. Plaintiff John Hornor Jacobs is an author who lives in Arkansas. He owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "Hornor Jacobs Asserted Works"): *The Conformity* (USCO Registration Number TX0008114163); *The Sea Dreams It Is the Sky* (USCO Registration Number TX0008830088); *The Shibboleth* (USCO Registration Number TX0007905434); ~~and~~ *This Dark Earth* (USCO Registration Number TX0007561564); *Twelve-Fingered Boy* (USCO Registration Number TX0009572498); *A Lush and Seething Hell* (USCO Registration Number TX0008804502); and *My Heart Struck Sorrow* (USCO Registration Number TX0008809460).~~.~~

~~13.~~16. Plaintiff Eboni ~~McKinnon~~ Mckinnon Smith (a/k/a Kianna Alexander) is an author who lives in North Carolina. She owns registered copyrights in the following works that are within the datasets that Apple reproduced without permission (collectively, the "~~McKinnon~~ Mckinnon Smith Asserted Works"): *Love for All Time* (USCO Registration Number TX0008672549); *Couldn't Ask for More* (USCO Registration Number TX0008668191); ~~and~~ *Tempo of Love* (USCO Registration Number TX0008448578) (re-released in 2023 as *Stardust*); *Back to Your Love* (USCO Registration Number TX0009170932); *This Tender Melody* (USCO Registration Number TX0008193076); *Every Beat of My Heart* (USCO Registration Number TX0008217057); *After Hours Redemption* (USCO Registration Number TX0009021729); *A Love Like This* (USCO Registration Number TX0008563355); *After Hours Attractions* (USCO Registration Number TX0009100320); *After Hours Temptations* (USCO Registration Number TX000922722); Then Came You (USCO Registration Number TX0008787155); *Carolina Built* (USCO Registration Number TX0009135026); *After Hours Agenda* (USCO Registration Number TX0009297307); *Love for All Time* (USCO Registration Number TX0008672549).

~~14.~~17. Below, the Hendrix Asserted Works, Roberson Asserted Works, Hornor Jacobs Asserted Works, and ~~McKinnon~~ Mckinnon Smith Asserted Works are referred to collectively as the "Infringed Works." On information and belief, Apple may have also infringed other works by Plaintiffs. A complete list of all of the Plaintiffs' registered copyrighted works is attached as Exhibit A to this Complaint.

B. Defendant

~~15.~~18. Defendant Apple Inc. is a California corporation with its principal place of business at One Apple Park Way, Cupertino CA 95014.

C. Agents and co-conspirators

~~16.~~19. The unlawful acts alleged against Apple in this class action complaint were authorized, ordered, or performed by Apple's respective officers, agents, employees, representatives, or shareholders while actively engaged in the management, direction, or control of Apple's businesses or affairs. Apple's agents operated under the explicit and apparent authority of their principals. Apple and its subsidiaries, affiliates, and agents operated as a single unified entity.

~~17.~~20. Various persons or firms not named as defendants may have participated as co-conspirators in the violations alleged herein and may have performed acts and made statements in furtherance thereof. Each acted as the principal, agent, or joint venture of Apple with respect to the acts, violations, and common course of conduct alleged herein.

V. FACTUAL ALLEGATIONS

~~18.~~21. Apple, currently the second largest company in the world by market capitalization, is an electronics and media company that designs, manufactures, and sells software and hardware technology products. Every second, Apple sells seven iPhones. Apple owns and operates the App Store, the primary platform for app developers to offer, and users to download, purchase, and subscribe to, mobile applications that run on Apple's mobile operating system, iOS. The App Store earns billions of dollars of revenue for Apple each year.

~~19.~~22. In or around April 2024, Apple released a suite of open-source generative-AI models called OpenELM. Just a few months later, in or around June 2024, Apple announced the development of its commercial artificial-intelligence platform, called Apple Intelligence. Apple Intelligence includes multiple generative-AI models and related tools and technologies. The day after Apple officially introduced Apple Intelligence—what one reporter likened to ushering in a "new era"—the company gained more than $200 billion in value: "the single most lucrative day in the history of the company."

~~20.~~23. In January 2025, the company reported its "best quarter ever" with revenue of $124.3 billion, twice citing Apple Intelligence in its press release announcing the same and deeming it a part of the company's "best-ever lineup of products and services." The technology is integrated across Apple's products, including iPhones, and is intended to "make[] apps and experiences even better and more personal." Analysts contend that Apple Intelligence could add $4 trillion to the company's market capitalization.

~~21.~~24. To train its generative-AI models, including the models that are part of Apple Intelligence, Apple first amassed an enormous library of data. Part of Apple's data library includes copyrighted works—including books created by Plaintiffs—that were copied without author or owner consent, credit, or compensation. In essence, Apple has created its own internal shadow library for the purpose of training current and future LLMs that are and will be integrated into Apple software, Apple apps, and third-party apps that are offered on the Apple App Store and run on Apple hardware.

~~22.~~25. Apple has not attempted to pay these authors and owners for their contributions to this potentially lucrative venture. Apple did not seek the licenses necessary to copy and use the copyrighted books provided to its models. Instead, it intentionally evaded payment by multiple methods: using books already compiled in pirated datasets, scraping books off the web, contracting for books from third parties, and scanning and copying books. Nor did Apple pay these authors and owners to make copies of their works for other purposes, beyond training.

A. How large language models work.

~~23.~~26. Artificial intelligence denotes software that is designed to algorithmically create an illusion of human reasoning or inference, often using statistical and mathematical methods. While called "artificial intelligence," these systems largely derive their value from human intelligence, as embodied within (among other things) the massive number of copyrighted works that are used to train them.

~~24.~~27. As noted above, though an LLM is a software program, it is not created the way most software programs are—by human software programmers writing code. Rather, LLMs like Apple's are trained by copying an enormous quantity of textual works and then feeding these copies into the

model. Typically, an LLM developer like Apple starts by identifying and copying its initial raw datasets containing such works. Then, these raw datasets are filtered and cleaned before they are used as model training data. For example, model developers often create software filters to remove certain categories of data that are undesirable for training, such as personally identifying information, copyright and license notices, or profanity and unsafe material. During this process, LLM developers often create derivative datasets that are subsets or altered versions of the original raw datasets. Importantly, because these additional derivative datasets also contain copies of parts of or entire textual works from the original raw datasets, including copyrighted works, the process of filtering and cleaning raw datasets to create usable training data for an LLM necessarily requires the creation of additional copies of the textual works in the raw datasets. Furthermore, because this gathering and processing of data is considered a strategic investment, the model developer often retains the datasets in a private data library, to remain available for future research and reuse.

~~25.~~28. The corpus of text ultimately used to train a model is called the "training dataset." Datasets may have multiple uses during the development process of an LLM—for instance, checking the performance of a model at intermediate stages of training, in an iterative development process. Therefore, even if a dataset is not part of a model's final training dataset, it may still be used in a model's development process and retained in a data library for use in development of a future model.

~~26.~~29. During training, the LLM copies and ingests each textual work in the training dataset and extracts protected expression from it. The LLM progressively adjusts its output to more closely approximate the protected expression copied from the training dataset. The LLM records the results of this process in a large set of numbers called "weights" that are stored within the model. These weights are entirely and uniquely derived from the protected expression in the training dataset. Generally, the more data the LLM copies during training, the better the LLM's ability to simulate the protected expression within that data as part of the LLM's output. Moreover, generally speaking, the more diverse the training data is in prose, format, or subject matter, and the higher the quality of the written works that are within the training data, the better the LLM will be at producing high quality textual

output. As a result, high-quality textual works, including longform creative works, are highly valuable to companies seeking to compile AI training datasets.

~~27.~~30. Once the LLM has copied and ingested the textual works in the training dataset and converted the protected expression into stored weights, the LLM can emit convincing simulations of natural written language in response to user prompts. Whenever an LLM generates a response to a user prompt, it is performing a computation that relies on these stored weights, with the goal of imitating the protected expression ingested from the training dataset.

~~28.~~31. A significant portion of the material in the datasets that Apple downloaded and used for the purposes of training its LLMs consists of copyrighted works—including books written by Plaintiffs and Class Members—that Apple copied without consent and without providing credit or compensation.

B. The OpenELM language models were trained on copyrighted works.

~~29.~~32. In April 2024, Apple first announced the availability of the OpenELM language models on its website: "[W]e release OpenELM, a state-of-the-art open language model. OpenELM uses a layer-wise scaling strategy to efficiently allocate parameters within each layer of the transformer model, leading to enhanced accuracy."

~~30.~~33. The set of OpenELM language models released in April 2024 included variants called OpenELM-270M, OpenELM-450M, OpenELM-1_1B, and OpenELM-3B. The main difference between these models is the parameter size; a larger parameter size means the model can store more weights and perform more complex tasks (requiring more computing power). For instance, Apple's OpenELM-3B language model is so named because the model stores three billion ("3B") weights derived from protected expression found in its training dataset.

~~31.~~34. Each OpenELM model is hosted on a website called Hugging Face, where it has a "model card," a file accompanying an AI model that typically describes the model, its intended uses and limitations, its training parameters, and the training dataset used to train the model. The model card for each OpenELM model states: "Our pre-training dataset contains RefinedWeb, deduplicated PILE, a subset of RedPajama, and a subset of Dolma v1.6, totaling approximately 1.8 trillion tokens."

~~32.~~35. Apple's GitHub repository confirms that "OpenELM was pretrained on public datasets. Specifically, our pre-training dataset contains RefinedWeb, PILE, a subset of RedPajama, and a subset of Dolma v1.6."

~~33.~~36. Apple also published a paper about OpenELM ("OpenELM Paper"). In a table called "Dataset used for pre-training OpenELM," shown at right, Apple reveals that a large quantity of training data comes from the "Books" subset of a dataset called "RedPajama." The OpenELM Paper does not further describe the contents of the RedPajama dataset.

Source	Subset	Tokens
RefinedWeb		665 B
RedPajama	Github	59 B
	Books	26 B
	ArXiv	28 B
	Wikipedia	24 B
	StackExchange	20 B
	C4	175 B
PILE		207 B
Dolma	The Stack	411 B
	Reddit	89 B
	PeS2o	70 B
	Project Gutenberg	6 B
	Wikipedia + Wikibooks	4.3 B

Table 2. Dataset used for pre-training OpenELM.

~~34.~~37. But information about the RedPajama dataset is available elsewhere. The RedPajama dataset is hosted on Hugging Face. According to the documentation for the RedPajama dataset that was available there until around April 2024, its "Books" component is a copy of the "Books3 dataset" that is "downloaded from Huggingface [sic]" when a user runs the script that automatically assembles the RedPajama dataset. Therefore, anyone who used the "Books" subset of the RedPajama dataset for training an AI model used a copy of the Books3 dataset. The documentation for the RedPajama dataset does not further describe the contents of Books3.

~~35.~~38. A paper describing the RedPajama dataset (the "RedPajama Paper") further states that Apple's OpenELM models were trained on RedPajama-V1, which was described as "a publicly available, fully open, best-effort reproduction of the training data . . . used to train the first iteration of LLaMA family of models." The RedPajama Paper's recitation of the components of the LLaMA training dataset includes six of the same components (with the same token counts) that are listed in Apple's paper on OpenELM: GitHub, Books, ArXiv, Wikipedia, StackExchange, and C4. The RedPajama Paper later confirms that the "Books" component, with 26B tokens, included Books3.

~~36.~~39. A description of the contents of the Books3 datasets is available elsewhere. Books3 is a component of a separate AI training dataset called the Pile that was curated by a research organization

called EleutherAI. In December 2020, EleutherAI introduced this dataset in a paper called "The Pile: An 800GB Dataset of Diverse Text for Language Modeling" ("The Pile Paper"). This paper describes the contents of Books3:

> Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker … Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset … We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.

~~37.~~40. Similarly, Shawn Presser, who compiled the Books3 dataset, publicly confirmed in 2020 that that dataset is composed of "all of Bibliotik" and includes .txt files containing the full text of approximately 196,640 books.

~~38.~~41. Bibliotik is one of several notorious "shadow library" websites. Others include Library Genesis (aka LibGen, Z-Library, or B-ok), Sci-Hub, and Anna's Archive. The AI-training community has long been interested in these shadow libraries because they host and distribute vast quantities of unlicensed copyrighted material. For that reason, these shadow libraries violate the U.S. Copyright Act.

~~39.~~42. The 196,640 books in the Books3 dataset exist in .txt file format. A .txt file (pronounced a "text" file) is a simple file format that stores text data without any formatting, fonts, or images. Accordingly, the Books3 dataset consists of the text of the underlying 196,640 books.

~~40.~~43. By using the entire text of each book in Books3, Apple used copies of entire works to train its OpenELM models.

~~41.~~44. Plaintiffs' Infringed Works are among the works in the Books3 dataset.

~~42.~~45. Until October 2023, the Books3 dataset was available from Hugging Face. At that time, the Books3 dataset was removed with a message that it "is defunct and no longer accessible due to reported copyright infringement."

~~43.~~46. Presser himself has acknowledged that "we almost didn't release the data sets at all because of copyright concerns."

~~44.~~47. Before October 2023, anyone who used the "Books" subset of the RedPajama dataset for training necessarily made a copy of the Books3 dataset. Based on the information revealed in the OpenELM research paper and on its model card on Hugging Face, this includes Apple.

~~45.~~48. In sum, Apple has admitted training its OpenELM large language models on a copy of the "Books" subset of the RedPajama dataset, which in turn is a copy of the Books3 dataset. Therefore, Apple trained its OpenELM models on a copy of Books3, a known body of pirated books.

~~46.~~49. Because Plaintiffs' Infringed Works are part of Books3, Apple trained OpenELM on one or more copies of the Infringed Works and directly infringed Plaintiffs' copyrights along with the copyrights of the Class.

C. The Apple Intelligence Foundation Language Models were trained on copyrighted works.

~~47.~~50. In June 2024, Apple announced its commercial AI software platform, called Apple Intelligence. Apple Intelligence includes several AI models developed by Apple. Several of these models are called the *Apple Intelligence Foundation Language Models*.

~~48.~~51. The Foundation Language Models were described in a paper of the same name released by Apple on July 29, 2024 (the "FLM Paper"). According to that paper, the models are multimodal, meaning they can process, understand, and generate multiple types of data, including text and images. It follows that the Foundation Language Models were trained on image datasets as well as sets of text data.

~~49.~~52. The adjective *foundation* is commonly used to describe AI models that have broad capabilities to perform a wide variety of tasks. Consistent with this, Apple describes its Foundation Language Models as "highly capable in tasks like language understanding, instruction following, reasoning, writing, and tool use … These foundation models are at the heart of Apple Intelligence."

~~50.~~53. The FLM Paper emphasizes the special importance of a foundation model's capacity to write: "[w]riting is one of the most critical abilities for large language models to have, as it empowers various downstream use[s]."

~~51.~~54. In the FLM Paper, Apple identifies two separate foundation language models: *AFM-server* and *AFM-on-device.* The AFM-server model is a larger model that is intended for use

through an Apple-operated cloud service called Private Cloud Compute. The AFM-on-device model, by contrast, is intended to be small enough to be used directly on Apple devices (e.g., iPhones and laptops). According to the FLM Paper, the AFM-on-device model is "initialize[d] … from a pruned 6.4B model (trained from scratch **using the same recipe as AFM-server**.)" (emphasis added). This means that both the AFM-server and AFM-on-device models are trained on the same corpus of training data.

~~52.~~55. In the FLM Paper, Apple reveals three sources of training data: "data we have licensed from publishers, curated publicly available or open-sourced datasets, and publicly available information crawled by our web-crawler, Applebot."

~~53.~~56. In describing the first source of training data—"data we have licensed"—Apple says only that it "identif[ied] and license[d] a **limited amount** of high-quality data from publishers" (emphasis added). In addition to being comparatively "limited" in quantity, Apple does not use this licensed data during the main phase of training the Foundation Language Models—what Apple calls "core pre-training"—but during a subsequent phase called "continued pre-training."

~~54.~~57. As to its second source of training data—"publicly-available or open-sourced datasets"—Apple does not elaborate on the specific datasets used, saying only, "We evaluated and selected a number of high-quality publicly-available datasets with licenses that permit use for training language models." Apple then filtered the datasets to remove personally identifiable information, copyright management information, and other undesired text before including them in the pre-training mixture.

~~55.~~58. In the parlance of AI training datasets, Apple's phrase "publicly available" is one commonly used to falsely conjure up the idea of works made publicly available by the author. But the "public" nature of a dataset does not mean that the data collected in the dataset was obtained lawfully or that the party providing copies of the dataset has authority to extend a valid license to use the underlying copyrighted works. When companies describing AI training datasets use the phrase "publicly available" to describe those datasets, it has in practice meant only that the datasets were downloaded somewhere from the public internet, which contains a vast number of copyrighted works

by authors and owners who have not granted a license for reproduction. There is a name for this kind of downloading of copies: copyright infringement. There is also a name for the infringing copies: pirated works.

~~56.~~59. For instance, Apple described its training data for OpenELM, including data from the Pile, as "public datasets," even though the Pile contains an extraordinary number of pirated copies of copyrighted works. As another example, Meta Platforms also trained its LLaMA language models on Books3, which it described as a "publicly available dataset for training large language models" despite the fact that none of the authors or owners whose works appear in Books3 ever consented to having their works included. Books3 was "publicly available" only in the limited sense that at one time, it could be acquired by anyone with an internet connection.

~~57.~~60. Similarly, in the context of AI training datasets, Apple's phrase "open source" is commonly used to falsely conjure up the idea of works made available by the author under a permissive copyright license (e.g., a Creative Commons license). In practice, what it really means is copies of works made freely available by someone other than the author, without the author's permission. Again, these are just pirated works.

~~58.~~61. For instance, EleutherAI, the group that created The Pile—the dataset that included Books3—described it as a "diverse, open source language modelling data set" even though the copyrighted works in the Books3 portion were included without authors' and owners' consent. Only the copyright holder can offer their copyrighted work to the public under an open-source license. A third party cannot usurp that right.

62. Therefore, when Apple says that a major source of the training data for its Foundation Language Models is "publicly-available or open-sourced datasets," this should, on information and belief, be read to mean "certain pirated works or certain other pirated works." Because Books3 has been described by people in the AI industry as a "publicly available" or "open-sourced" dataset, and because Apple already had a copy of Books3 that it had used for training its OpenELM models, on information and belief, Apple's reference to "publicly-available or open-sourced datasets" includes Books3, and Apple therefore included Books3 in its private data library, which formed the basis for the

training dataset for its Foundation Language Models. Thus, it is also, on information and belief, the case that the "curated publicly available or open-sourced datasets" that Apple copied to create a private data library for the purposes of training its Foundation Language Models (and future models) contain copyrighted material, including Plaintiffs' and Class Members' copyrighted works.

63. Plaintiffs' Infringed Works are part of Books3. It follows that, on information and belief, Apple trained its Foundation Language Models on one or more copies of the Infringed Works, thereby directly infringing the copyrights of the Plaintiffs.

64. On information and belief, Apple has created a permanent ~~AI training data~~digital library—and likely multiple such libraries—containing copies of all of its "curated," purportedly "publicly-available or open-sourced" datasets, like Books3, in expectation of future uses ~~training future models~~.

~~59.~~65. In obtaining a copy or copies of Books3, Apple made unauthorized reproductions of Plaintiffs' Infringed Works.

~~60.~~66. As to ~~its~~ a third source of training data—web pages crawled by Applebot—Apple says, "we crawl publicly available information using our web crawler, Applebot … and respect the rights of web publishers to opt out of Applebot." While Apple officially acknowledged Applebot's crawling of webpages in mid-2015, its activity had been spotted in logs as early as November 2014. Around June 2024, Apple revealed that it was using Applebot-scraped data for training its AI models. In response to this disclosure, by August 2024, numerous major commercial web publishers had chosen to opt out of Applebot training.

~~61.~~67. But Apple's Foundation Language Models had necessarily been trained well before the release of the FLM Paper describing them in July 2024. For that reason, Apple's disclosure in June 2024 that it was using Applebot data to train language models came too late for any of these opt-outs to matter. Apple had already scraped the data and trained language models with it. On information and belief, Apple has retained copies of all Applebot data scraped before this wave of opt-outs, in expectation of training future models, as part of its AI training-data library.

~~62.~~68. In a November 2024 paper by George Wukoson and Joey Fortuna called "The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLMs," the authors studied LLM training datasets made by algorithmically filtering scraped web pages. The authors concluded that such "datasets are disproportionately composed of high-quality content owned by commercial publishers of news and media websites." In turn, this material is often covered by registered copyrights. Thus, on information and belief, the parts of Apple's private shadow library and training datasets that come from filtered Applebot pages includes copyrighted works from commercial news and media websites.

~~63.~~69. The shadow libraries that host millions of unlicensed copyrighted books are also part of the "publicly available information" reachable by a web scraper like Applebot. Hence, on information and belief, part of Apple's training-data library is sourced from shadow libraries via Applebot.

On information and belief, Apple obscures the libraries and datasets used to train its Foundation Language Models to conceal its use of copyrighted materials. On information and belief, Apple's decision not to disclose the contents of its internal data libraries, or the training datasets for its Foundation Language Models, stems in part from the fact that Apple was the subject of negative press for using a subset of data from The Pile, containing captions from thousands of YouTube videos, at least in training the OpenELM models and, on information and belief, also in training the Foundation Language Models.

70. In obtaining copies of Plaintiffs' Infringed Works from web scrapes via Applebot, Apple made additional unauthorized reproductions of Plaintiffs' Infringed Works.

71. Apple's sources of training data also included purchased print copies of books that were scanned into Apple's digital systems; Apple's existing troves of books and audiobooks, including within Apple Books; and third parties' data archives, which Apple contracted to access. Like the pirated and scraped datasets, at least some of the texts in each of these categories and sources, too, were copied after acquisition and used for training as well as other purposes, including the creation of permanent data libraries, the performance of research, and the transfer of copies to third parties.

~~64.~~72. As part of its development of Apple Foundation Models (AFMs), also known as Foundation Language Models (FLMs), Apple created "teacher" models that generated synthetic data, "classifier" models that filtered models' inputs and outputs, applications based on fine-tuned versions of those models, and multiple generations of the models themselves, including the third generation released in June 2026. On information and belief, the training data for these models and applications includes data from Apple's private data libraries, which encompass Plaintiffs' Infringed Works through datasets including, but not limited to, Books3 and Library Genesis.

73. Most recently, in June 2026, Apple publicly released to developers the third generation of its language models, now called Apple Foundational Models or "AFM." This update includes four new language models, called "AFM 3 Core," "AFM 3 Core Advanced," "AFM 3 Cloud," and "AFM 3 Cloud Pro."

74. In a press release, Apple describes the training data for its AFM 3 models: "To train our foundation models, we use a mixture of data that includes publicly available information, data licensed or purchased from third parties, open-sourced data, data obtained through dedicated studies, and synthetic data." These categories are essentially equivalent to the categories of training data disclosed in the FLM paper (see ¶ 55) with two new ones: "data obtained through dedicated studies" and "synthetic data." On information and belief, "synthetic data" refers to data generated by other Apple language models, including models that were trained on datasets including copies of Plaintiffs' Infringed Works such as Books3 and Library Genesis.

75. Apple executives like Craig Federighi and Amar Subramanya have stated that the third-generation of Apple Foundation Models were the product of Apple's collaboration with Google and were trained using proprietary data and refined using outputs from Gemini frontier models. On information and belief, these statements that Apple's new models are "trained using proprietary data" means that they were trained using data, including Plaintiffs' Infringed Works, that Apple had stored in its private data library after training the earlier Foundation Language Models.

D. Apple's models for data classification were trained on copyrighted works.

~~65.~~76. In the FLM Paper, Apple says that Applebot pages are "processed by a pipeline which performs quality filtering … using heuristics and model-based classifiers." In this context, the term "model-based classifier" refers to a separate AI model that has been trained to algorithmically rate the quality of and sort scraped web pages. On information and belief, the models that underlie these model-based classifiers were themselves trained on datasets that include unlicensed copyrighted works, including Plaintiffs' Infringed Works.

E. All of Apple's LLMs and model-based classifiers likely used multiple shadow libraries.

~~66.~~77. As copyright-infringement litigation has proceeded against other companies training LLMs, it has become clear that these companies downloaded, torrented, and/or used more pirated works and shadow libraries than was publicly known at the time of the complaint. Across these suits, the additional shadow libraries implicated include Anna's Archive, Pirate Library Mirror ("PiLiMi"), Z-Library (also known as B-ok), LibGen, and Sci-Hub. *See, e.g.*, Joint Letter Brief, *In re OpenAI ChatGPT Litig.*, No. 3:23-cv-03223-AMO (N.D. Cal. Jan 17, 2025) (Dkt. No. 254, at 1, 5) (acknowledging OpenAI's production of datasets "LibGen 1" and "LibGen 2"); First Consolidated Amended Complaint, *Nazemian, et al. v. NVIDIA*, No. 4:24-cv-01454-JST (N.D. Cal. Jan. 16, 2026) (Dkt. No. 235, at 12-13) (describing communications between NVIDIA and Anna's Archive, and alleging pirating from LibGen, Z-Library, Pirate Library Mirror, and Sci-Hub); Order Denying the Plaintiffs' Motion for Partial Summary Judgment and Granting Meta's Cross-Motion for Partial Summary Judgment, *Kadrey, et al. v. Meta*, No. 3:23-cv-03417-VC (N.D. Cal. June 25, 2025) (Dkt. No. 598, at 11) ("In early 2024, Meta also downloaded Anna's Archive, a compilation of shadow libraries including LibGen, Z-Library, and others."); Order on Fair Use, *Bartz, et al. v. Anthropic*, No. 3:24-cv-05417-WHA (N.D. Cal. June 23, 2025) (Dkt. No. 231, at 3) (noting that Anthropic downloaded millions of copies of books from LibGen and PiLiMi). According to the administrators of Anna's Archive, "virtually all major companies building LLMs contacted us to train on our data. . . We have given high-speed access to about 30 companies." The piracy is so widespread as to constitute a *de facto* industry practice.

~~67.~~ Apple, like the companies listed above, is also a major company competing to build LLMs. As there was with the companies above, there is publicly available information indicating that Apple downloaded copyrighted works and used them in the multiple ways alleged herein. Thus, on information and belief, Apple, like the companies listed above, also very likely accessed one or more of these notorious shadow libraries to obtain large volumes of illegal copies of copyrighted books. Like other major technology companies that did the same thing, Apple, on information and belief, did so while avoiding paying the customary fee for books and without permission from the copyright holder.

78. Apple's internal documents reveal that Apple's Artificial Intelligence Machine Learning ("AIML") Team used Books3 to train many, and perhaps all, of its large language models. In December 2023, an Apple engineer noted that "essentially all models" that Apple had created ~~to date~~as of that time were trained on Books3.

79. Other internal documents cite to "libgen," which is shorthand for the shadow library Library Genesis. Like Books3, Library Genesis is made up of unlicensed copyrighted material. Library Genesis currently contains more than 7.5 million books and 81 million research papers.

80. In June 2023, Apple employees suggested that while they could not retrieve contents from Library Genesis "directly," they could likely find a "direct dump" of the contents on another website like Huggingface, where Apple employees sourced a copy of Books3.

81. In July 2025, Apple employees stated: "[s]ince our current research text cont pretrain data contains libgen, we have to regenerate one without libgen"; and "out [sic] current continued pretrain mixture includes libgen." Other employees responded "Yeah, unfortunately, we need to remove that for now," and "[w]e can train on it internally but can not release any models including that."

82. By downloading a copy or copies of Library Genesis, Apple made unauthorized copies of Plaintiffs' Infringed Works. These unauthorized copies constituted copyright infringement regardless of what, if anything, Apple later did with the books.

F. Apple integrates the Foundation Language Models, including via Apple Intelligence, into Apple's commercial products.

~~68.~~83. Apple has integrated its Apple Intelligence products, including its Foundation Language Models, into a wide variety of commercially available products, including versions of the iPhone 15, 16, and 17; the iPad, iPad mini, and iPad Pro; the MacBook Air, MacBook Pro, iMac, Mac mini, and Mac Studio; and Apple Vision Pro.

~~69.~~84. In addition, Apple has offered Apple Intelligence, and the FLM in particular, to developers of Apple device software applications, or "apps," as a way to "create new intelligence features" using "the on-device large language model" at the core of Apple Intelligence. Apps are software applications, designed to run on devices such as smartphones, tablets, and other devices, that permit end-users to accomplish certain tasks. With many apps available on Apple devices, the end-users are typically consumers. According to Apple's public-facing site for app developers, the "Foundation Models framework provides" developers "access to Apple's on-device large language model that powers Apple Intelligence," and can be used by developers to create apps that permit consumers to "generat[e] creative content" and complete "a diverse range of text generation tasks." Apple's site specifically highlights the ability of Apple's FLM on-device model to "[c]ompose creative writing," even including, as an example, the following prompt an app user could enter: "Generate a short bedtime story about a fox." Additionally, Apple has released code and a machine-learning framework to assist development of iOS apps that integrate the OpenELM models.

85. Unsurprisingly, within several months of the Foundation Language Models becoming available to Apple developers for iOS 26, apps have appeared in the Apple App Store that allow users to generate output that could, in the short- or long-term, compete with copyrighted works. For example, Apple touted the app "Lil Artist" as one app relying on FLM, describing it as "combin[ing] the capabilities of the Foundation Models framework and the ImageCreator API to customize illustrated stories for children." Another app available through the Apple App Store, called "Locally," includes a "write professionally" button that encourages creative, long-form content generation. These apps, similar apps that currently exist, and similar apps that will inevitably be created in the near future

using the FLM models or an OpenELM model, can be used to generate output that could compete with, dilute the market for, or simply be perceived and consumed by consumers as substitutes for Plaintiffs' and Class Members' copyrighted works. Apple earns revenue from the distribution and use of these apps via its App Store.

~~70.~~86. In September 2026, the AFM 3 models were released to the public as part of Apple's OS 27 update.

~~71.~~87. Apple has reportedly explored a paid tier for users of its Apple Intelligence products, with some early reports suggesting it may do so as early as 2027.

G. Apple scanned copies of Plaintiffs' and Class Members' books and then made numerous subsequent copies for training and reuse in a central repository.

88. In 2025, Apple decided to~~that~~ begin to scan books as a method to obtain more textual data to use for trainings~~s~~ its models. Apple then launched "Project Borders," its project focused on digitizing books to use in trainings~~s~~ its AFM models. In July 2025, Apple noted it had around [redacted] books to scan. Apple then procured vendors to obtain and scan books, which provided Apple with digitized copies of books.

89. Apple decided to expand its book--scanning project after determining it was cheaper to buy, scan, and digitize books than to license copies of eBooks from publishers.

90. By October 2025, Apple was scanning [redacted] books ~~a~~per day and had a goal of scanning [redacted] books by the end of March~~,~~ 2026. On information and belief, Apple made additional copies of the books that it scanned, some of which were used for training, some of which were used for, and stored in Apple's central repositories for future uses including, research and transfer to third parties.

H. Apple built and still maintains central repositories of Plaintiffs' and Class Members' works for future uses.

91. While Apple sources massive amounts of data for training its large language models, it does not delete such data after training its models. Instead, Apple retains these datasets in private data repositories that remain available for future research and reuse.

92. Apple's internal documents show that Apple has amassed central repositories of electronic versions of books and other printed material from a variety of sources, including (1) book scanning; (2) book/journal licensing; (3) Internet Archive downloads; and (4) open web PDF downloads. As one Apple engineer stated, "We are really building Apple's repository of all the printed information in the world."

93. Documents describing Apple's process for scanning books indicate that process, in which vendors send scanned PDF versions of books to Apple, Apple validates and processes the scanned book data, and that data "is then centrally stored."

94. Apple uses a variety of methods, vendors, and processes for storing its data in central repositories. Apple's repositories of electronic versions of books include, but are not limited to, data buckets (repositories hosted by various third-party services) on [redacted], Amazon S3, Google Cloud Storage. Apple also maintains data repositories for its foundational model data and synthetic data on Github.

95. Apple makes its central repositories accessible across teams, including those other than the AIML team responsible for designing and training Apple's AFM models. Internal documents show that Apple planned for all AFM datasets, models, and lineage information (i.e., where the data came from) to be "easily accessible" by various Apple teams, including AIML, Apple Services Engineering ("ASE"), Video Engineering ("VE"), and Software Engineering ("SWE").

96. Even when Apple identifies a dataset of textual works as being problematic for training publicly released large language models, Apple still maintains copies of those datasets in its central repositories for future research and additional uses.

97. On information and belief, Apple maintains copies of Plaintiffs' Infringed Works and Class Members' works in its central repositories for future uses.

~~G.~~I. **Apple's conduct impairs the markets for Plaintiffs' and Class Members' works.**

98.

~~72.~~ Apple has neither paid nor sought permission from Plaintiffs for the use of their copyrighted works. Instead of doing so, Apple downloaded, torrented, scraped, or otherwise copied

vast quantities of copyrighted works—including illegally compiled datasets such as Books3—that included Plaintiffs' works like the Infringed Works.

~~73.~~99. In so doing, Apple has, first, deprived Plaintiffs of the revenue that would have been generated had Apple approached Plaintiffs or their licensing agents directly to license copies of their works. Plaintiffs and similarly situated creators previously licensed their work for their own commercial uses.

~~74.~~100. There are numerous examples of publicly reported AI licensing deals. For example, Microsoft reportedly reached an agreement with HarperCollins to license a select number of non-fiction titles to train AI models for just three years; compensation for participating authors, who have to opt-in before their works can be licensed, is $2,500 per work. Wiley likewise reported in 2025 that it had executed multiple AI licensing agreements totaling $40 million. Myriad licensing systems have been launched and are continuing to develop, including the Copyright Clearance Center's collective AI licensing scheme and the Created by Humans licensing platform. Further, several AI dataset licensing companies have formed a trade group called the Dataset Providers Alliance in order to promote standardization and transparency in the licensing of intellectual property content for AI and machine learning datasets. Currently, some researchers estimate, the AI training license market is valued at approximately $2.5 billion; within a decade, it may close in on nearly $30 billion.

~~75.~~101. Apple itself understands the value of copyrighted works and the market that exists for paying creators to use their works for training. For instance, it struck an agreement with Shutterstock to "use hundreds of millions of images, videos and music files" valued between an estimated $25 to $50 million. Similarly, Apple has contacted news organizations like Condé Nast, NBC News, and IAC to license news article archives. Nonetheless, Apple has not compensated Plaintiffs' and Class Members whose works it copied and used in training its models.

~~76.~~102. Apple has not only directly deprived Plaintiffs and Class Members of the licensing revenues Plaintiffs and Class Members could have earned from Apple for Apple's use of their works to train their LLMs. Apple has also contributed, through its conduct, to the impairment of the emergence of lawful licensing regimes. Such nascent licensing regimes are more successful if large

corporate licensees participate early and at scale, contributing significant licensing revenues and stabilizing the demand side of such a licensing market. Apple's choice to instead copy pirated works on a massive scale has, merely through the absence of a corporate cornerstone of such a market, blunted the development of an AI training dataset licensing regime.

~~77.~~103. Apple's conduct has also depressed the demand for licensing from other potential licensees by normalizing and encouraging other potential licensees not to license works for AI training. Apple has significant consumer brand loyalty and is one of the most valuable companies in the world. Smaller potential licensees are less likely to pursue licensing if they see that Apple, with its cadre of retained and in-house lawyers, has determined it does not need to do so. Smaller potential licensees are also less likely to pursue licensing if they believe they cannot do so and remain competitive with larger companies with more capital (like Apple) who have declined to pay licensing fees and instead flouted copyright law. Finally, smaller potential licensees are less likely to pursue licensing if Apple's conduct has normalized the use of pirated works, which use Apple has largely attempted to hide from consumers.

~~78.~~104. Second, Apple's unauthorized use of Plaintiffs' copyrighted works to train the OpenELM models and Foundation Language Models has also caused and threatens to cause substantial harm to the actual and potential markets for those works. As described above, Apple has developed, designed, created, maintained, and/or distributed functionalities and applications that are designed to, and can, generate works that imitate or otherwise resemble the Infringed Works. Apple has also encouraged and facilitated third parties' design, creation, maintenance, and/or distribution of relevantly similar applications. All of these unauthorized uses of the Infringed Works creates a risk of market dilution, which may "lead to a loss of sales" by "harm[ing] the market for access to those works." Works generated using the OpenELM models or the Foundation Language Models will inevitably start competing with Plaintiffs' works (like the Infringed Works) and ultimately dilute royalty pools as AI-generated output increasingly floods the market. Already, "low-quality sham 'books'" created by other LLMs have begun overwhelming the market as scammers generate "unauthorized 'biographies' of authors that are simply AI-generated rehashings of their lives, often based on autobiographical works."

Other scams include "companion books" that summarize the key points from the original novel, with "little to no original analysis or commentary and are meant only to confuse consumers and skim sales off of the real books." These works have already entered book marketplaces like Amazon.

~~79.~~105. Apple's models—whether integrated in Apple Intelligence, an Apple-developed app, or a third-party-developed app—are designed to, and can, generate outputs that may substitute the kinds of expressive written work that Plaintiffs are hired to produce, potentially diminishing demand for books and human-produced stories. Plaintiffs face potential ongoing harms through lost publication opportunities, reduced recognition, and lost sales, among other harms.

~~80.~~106. This type of substitution of human-created works for AI-generated works has already been documented in creative markets. For example, the *New York Times* reported in a recent front-page story: "Publishers and authors worry that books by real writers are getting lost in the sea of digital slop, as A.I.-enabled novels flood the market." According to one best-selling novelist quoted in that article: "It bogs down the publishing ecosystem that we all rely on to make a living . . . It makes it difficult for newer authors to be discovered, because the swamp is teaming [sic] with crap." Although some stigma currently attaches to known AI-generated novels, another writer predicts, "[e]ventually . . . readers will not care." As an additional example, researchers have recently shown that some readers prefer the outputs of AI models trained on copyrighted works over the work of human writers.

~~81.~~107. Additionally, in a 2025 working paper analyzing the impact of generative AI on the stock image market, the authors describe their findings as follows:

> In this paper, we ask three questions with the aims of shedding light on how GenAI will change markets for creative goods and informing the active debate on copyright and fair use.
>
> **First, to what extent do GenAI goods compete with—or become substitutes for—non-GenAI goods? We provide causal evidence that consumers view GenAI-produced goods as substitutes for traditional creative goods. Thus GenAI goods are competitive with non-GenAI goods.** Further, the market expands as GenAI goods enter the market, with GenAI goods consumption more than replacing non-GenAI goods.
>
> **Next, does GenAI result in crowd-out of non-GenAI firms and goods?** Policymakers in particular have expressed this concern given the substantial differences in production costs brought by GenAI. **Our results show that GenAI does crowd out non-GenAI firms and goods. Across almost all markets, we see substantial entry of GenAI firms and exit of non-GenAI firms, with net production expansion, but a decrease**

> **in any given product's probability of sale. We document that this concern is particularly relevant for niche markets, where we find production tipping almost completely to GenAI.** . . .
>
> . . .
>
> The substitution to non-GenAI content, and significant exit of non-GenAI artists, provide evidence of crowd-out of non-GenAI content. **In the long run, this may lead to a decline in the production of novel original content for model training, and even to a decline in non-GenAI content altogether. Further, we document a reduction in sales rates for artists.** While these reductions may be partially compensated for by the increase in production, not all artists, and in particular not all non-GenAI artists, will be able to fully compensate via scale, and thus their long-run profitability may be at risk. **Thus, our results provide some validity to concerns that GenAI content will both replace non-GenAI content and greatly diminish the value of the original content, affecting the fourth pillar of fair use.**

VI. CLASS ACTION ALLEGATIONS

~~82.~~108. The "Class Period" as defined in this Complaint begins at least three years before September 5, 2025 and runs through the present.

~~83.~~109. As used here, the term "Apple Uses" refers collectively to Apple's numerous separate infringing uses of Plaintiffs' and Class Members' works, including the Infringed Works, as set forth below:

A. *Acquiring Data via Pirated Sources:* Apple's unauthorized reproduction of Plaintiffs' and Class Members' works, including the Infringed Works, as part of Apple's downloading, torrenting, or other mode of initial acquisition of one or more datasets of pirated works, including Books3, Library Genesis, Z-Library, Anna's Archive, Sci-Hub, Internet Archive, and/or Pirate Library Mirror;

B. *Acquiring Data via Applebot:* Apple's unauthorized copying of Plaintiffs' and Class Members' works, including the Infringed Works, as part of Apple's scraping of the public internet using Applebot or other methods;

~~B.~~ *Reproducing Data after Acquisition:* Apple's unauthorized copying of Plaintiffs' and Class Members' works, including the Infringed Works, whether for the purpose of training, research, transfer, or permanent retention, regardless of the mode of acquisition (including but not limited to downloading, scraping, licensing, contracting, and scanning);

C. *Curating:* Apple's unauthorized reproduction and use of Plaintiffs' and Class Members' works, including the Infringed Works, when filtering, cleaning, and processing at least one pirated dataset, such as Books3, and components of such datasets, to create a curated, private shadow library of datasets for use in various stages of training LLMs;

D. *Training OpenELM Models:* Apple's unauthorized reproduction and use of Plaintiffs' and Class Members' works, including the Infringed Works, when training its OpenELM models;

E. *Training FLMs:* Apple's unauthorized reproduction and use of Plaintiffs' and Class Members' works, including the Infringed Works, when training all versions and generations of its~~its~~ Foundation Language Models ~~and all versions and generations thereof~~;

F. *Training Classifier and Teacher Models:* Apple's unauthorized reproduction and use of unlicensed copyrighted works when training classifier and teacher models; and

~~G.~~ *Retaining for Private Data Library:* Apple's unauthorized retention, curation, reproduction, and use of Plaintiffs' and Class Members' works, including the Infringed Works, along with all the data Apple has illegally downloaded, ~~and~~ curated, reproduced, and used thus far, in the form of a private data library for use in models currently under development as well as future models—an AI data library that includes the Books3 dataset, which, in turn, includes the Infringed Works.

~~84.~~110. **Class definition.** Plaintiffs bring this action for damages and injunctive relief as a class action under Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3), on behalf of the following Class:

> **All beneficial or legal owners of a registered United States copyright in any work that was 1) used in one or more of the Apple Uses during the Class Period, and 2) was registered with the United States Copyright Office a) within five years of the work's publication and before its use in one or more of the Apple Uses, or b) within three months of publication.**

~~85.~~111. This Class definition excludes: (a) Defendant; (b) Any of Defendant's parent companies, subsidiaries, and affiliates; (c) Any of Defendant's officers, directors, management,

employees, subsidiaries, affiliates, or agents; (d) All governmental entities; and (e) The judges and chambers staff in this case, as well as any members of their immediate families.

~~86.~~112. **Numerosity: Federal Rule of Civil Procedure 23(a)(1).** The Class Members are so numerous and geographically dispersed that individual joinder of all Class Members is impracticable. The exact number of Class Members is currently unknown to Plaintiffs, as this information is in Defendant's exclusive control. On information and belief, there are more than several thousand members in the Class across the United States. Accordingly, joinder of all Class Members in prosecuting this action is impracticable.

~~87.~~113. The Class can be identified, in part, through tools that allow a user to search for web domains included in the RedPajama dataset, the Pile dataset, and other datasets used for one or more of the Apple Uses.

~~88.~~114. The Class can further be identified by obtaining business records maintained by Apple, including the content of its private pirated libraries and datasets used for both its OpenELM models and Foundation Language Models.

~~89.~~115. **Typicality: Federal Rule of Civil Procedure 23(a)(3).** Plaintiffs' claims are typical of the claims of Class Members because Plaintiffs and all members of the Class were damaged by the same course of conduct of Defendant. Further, the relief sought is common to all Class Members.

~~90.~~116. **Adequacy of Representation: Federal Rule of Civil Procedure 23(a)(4).** Plaintiffs will fairly and adequately represent the interests of the members of the Class because Plaintiffs have experienced the same harms as the members of the Class and have no conflicts with any other members of the Class. Further, Plaintiffs have retained competent counsel who are experienced in litigating federal class actions and other complex litigation involving sophisticated, state-of-the art technology.

~~91.~~117. **Commonality and Predominance: Federal Rules of Civil Procedure 23(a)(2) and 23(b)(3).** Numerous questions of law and fact are common to each Class Member arising from Defendant's conduct, including:

A. Whether Apple acquired or downloaded reproductions of datasets, including but not limited to the RedPajama and Books3 datasets, that included Plaintiffs' and Class Members' copyrighted works;

B. Whether Apple copied Plaintiffs' and Class Members' works as part of Apple's scraping of the public internet using Applebot or other methods;

C. Whether Apple included Plaintiffs' and Class Members' works in any of the datasets used by Apple to train its OpenELM models and Foundation Language Models;

D. Whether Apple's inclusion of Plaintiffs' and Class Members' works in those datasets constituted or required the works' reproduction by Apple;

E. Whether Apple lacked authorization to reproduce copies of Plaintiffs' and Class Members' works;

F. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and used them in training its OpenELM models;

G. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and used them in training its Foundation Language Models;

H. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and retained them in a curated, private shadow library for Apple's use in training other models;

I. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and reproduced them for purposes of research;

~~H.~~J. Whether Apple violated the copyrights of Plaintiffs and the Class when it acquired or downloaded copies of Plaintiffs' and Class Members' works and reproduced and transferred them to third parties;

~~I.~~K. Whether this Court should enjoin Defendant from engaging in the unlawful conduct alleged herein;

~~J.~~L. Whether any affirmative defense excuses Defendant's conduct, including the fair use doctrine; and

~~K.~~M. Whether Apple's infringement was willful.

~~92.~~118. These and other questions of law and fact are common to the Class and predominate over questions affecting Class Members on an individual basis.

~~93.~~119. **Predominance & Superiority: Federal Rule of Civil Procedure 23(b)(3).** Defendant has acted on grounds generally applicable to the Class. A class action is superior to alternatives for the fair and efficient resolution of this controversy. Allowing the claims to proceed on a class basis will eliminate the possibility of repetitive litigation. Further, injunctive relief is appropriate with respect to the entire Class.

~~94.~~120. **Risk of Prosecuting Separate Actions.** The alternative of separate actions by individual Class Members risks inconsistent adjudications and is an inefficient use of limited judicial resources.

VII. CLAIM

DIRECT COPYRIGHT INFRINGEMENT — 17 U.S.C. § 501

~~95.~~121. Plaintiffs incorporate by reference all other allegations in this complaint.

~~96.~~122. As the owners of the registered copyrights in the Infringed Works and other copyrighted works, Plaintiffs and Class Members hold the exclusive rights to those works under 17 U.S.C. § 106. Plaintiffs and the Class Members never authorized Apple to make copies of their Infringed Works and other copyrighted works, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works), or exploit any other right exclusively reserved to Plaintiffs and the Class Members under the U.S. Copyright Act.

~~97.~~123. In performing the Apple Uses, including by downloading the works and making additional copies during various stages of training and development of the OpenELM models and the

Foundation Language Models, Apple made all its copies of the Infringed Works and other copyrighted works without Plaintiffs' or Class Members' permission, violating their exclusive rights under the U.S. Copyright Act. Indeed, "the person who copies the textbook from a pirate site has infringed already, full stop." *Bartz et al. v. Anthropic*, 787 F. Supp. 3d 1007, 1025 (N.D. Cal. 2025). Regardless of how Apple uses the works in its private training-data library in the future, this cannot negate that the initial copying of works sourced from shadow libraries infringed on Plaintiffs' and Class Members' exclusive rights.

~~98.~~124. Plaintiffs and Class Members have been injured by Apple's acts of direct copyright infringement of the Infringed Works. Plaintiffs and Class Members are entitled to statutory damages, actual damages, restitution of profits, destruction of the infringing copies and models, and other remedies provided by law.

~~99.~~125. Apple's violation of Plaintiffs' and Class Members' exclusive rights was willful. Apple knew the datasets it downloaded, copied, stored, and trained its LLMs on contained copyrighted works.

VIII. PRAYER FOR RELIEF

~~100.~~126. Plaintiffs demand judgment on their behalf and on behalf of the Class against each Defendant as follows:

A. Allowing this action to proceed as a class action, with Plaintiffs serving as Class Representatives, and with Plaintiffs' counsel as Class Counsel;

B. Awarding Plaintiffs and the Class statutory damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity;

C. Permanently enjoining Defendant from the unlawful, unfair, and infringing conduct alleged herein;

D. Ordering destruction under 17 U.S.C. § 503(b) of all Apple LLMs that have ingested at least one copy of any of Plaintiffs' and Class Members' works;

E. Ordering destruction under 17 U.S.C. § 503(b) of all copies of those works that are maintained in Apple's private libraries and datasets;

F. An award of costs, expenses, and attorneys' fees as permitted by law; and

G. Such other or further relief as the Court may deem appropriate, just, and equitable.

IX. DEMAND FOR JURY TRIAL

Plaintiffs demand a jury trial for all claims.

DATED this 2nd day of November, 2026.

KELLER ROHRBACK L.L.P.

By *s/ William K. Dreher*

William K. Dreher, *Pro Hac Vice*
Benjamin Gould (SBN 250630)
Derek W. Loeser, *Pro Hac Vice*
~~Cari C. Laufenberg, *Pro Hac Vice*~~
Chris N. Ryder, *Pro Hac Vice*
~~Elizabeth W. Tarbell, *Pro Hac Vice*~~
~~Samuel Rubinstein, *Pro Hac Vice*~~
1201 Third Avenue, Suite 3400
Seattle, WA 98101-3268
(206) 623-1900
Fax (206) 623-3384
bgould@kellerrohrback.com
wdreher@kellerrohrback.com
dloeser@kellerrohrback.com
cryder@kellerrohrback.com
~~claufenberg@kellerrohrback.com~~
~~etarbell@kellerrohrback.com~~
~~srubinstein@kellerrohrback.com~~

SUSMAN GODFREY L.L.P.

By *s/ Rohit D. Nath*

Rohit D. Nath (SBN 316062)
Justin A. Nelson, *Pro Hac Vice*
Alejandra C. Salinas, *Pro Hac Vice*
Jordan Connors, *Pro Hac Vice*
Eleanor Runde, *Pro Hac Vice*

1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
(310) 789-3100
rnath@susmangodfrey.com
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com
jconnors@susmangodfrey.com
erunde@susmangodfrey.com

BUTTERICK LAW PC

By *s/ Matthew Butterick*
Matthew Butterick (SBN 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
mb@buttericklaw.com

Attorneys for Plaintiffs

4900-1942-3710, v. 1~~4915-4826-0826, v. 15~~4915-4826-0826, v. 1~~4~~4915-4826-0826, v. ~~13~~

Exhibit A

Exhibit A

Registration Number	Title of Work	Author of Work
TX0008563355	A Love Like This	Eboni Manning aka Kianna Alexander
TX0009297307	After Hours Agenda	Eboni Manning aka Kianna Alexander
TX0009100320	After Hours Attraction	Eboni Manning aka Kianna Alexander
TX0009021729	After Hours Redemption	Eboni Manning aka Kianna Alexander
TX0009227221	After Hours Temptation	Eboni Manning aka Kianna Alexander
TX0009170932	Back to Your Love	Eboni Manning aka Kianna Alexander
TX0009341367	Can’t Let Her Go	Eboni Manning aka Kianna Alexander
TX0009154431	Can’t Resist Her	Eboni Manning aka Kianna Alexander
TX0009135026	Carolina Built	Eboni Manning aka Kianna Alexander
TX0008668191	Couldn’t Ask for More	Eboni Manning aka Kianna Alexander
TX0008217057	Every Beat Of My Heart	Eboni Manning aka Kianna Alexander
TX0008744613	Forever With You	Eboni Manning aka Kianna Alexander
TX0008672549	Love for All Time	Eboni Manning aka Kianna Alexander
TX0008979954	Never Let Me Go	Eboni Manning aka Kianna Alexander
TX0008448578	Tempo Of Love	Eboni Manning aka Kianna Alexander
TX0008787155	Then Came You	Eboni Manning aka Kianna Alexander

Exhibit A

Registration Number	Title of Work	Author of Work
TX0008193076	This Tender Melody	Eboni Manning aka Kianna Alexander
TX0009242549	What Happens After Hours	Eboni Manning aka Kianna Alexander
TX0009327754	Ankle Snatcher	Grady Hendrix
TX0009281005	BadAsstronauts	Grady Hendrix
TX0007957438	Horrorstor: A Novel	Grady Hendrix
TX0009231440	How to Sell a Haunted House	Grady Hendrix
TX0008276604	My Best Friend's Exorcism	Grady Hendrix
TX0008461565	Paperbacks from Hell: The Twisted History of '70s and '80s Horror Fiction	Grady Hendrix
PAu002397892	Satan Loves You	Grady Hendrix
TX0009212410	The Final Girl Support Group	Grady Hendrix
TX0008888592 and TX0008986901	The Southern Book Club's Guide to Slaying Vampires: A Novel	Grady Hendrix
TX0008645317 and TX0008645012	We Sold Our Souls: a Novel	Grady Hendrix
TX0009505624	Witchcraft for Wayward Girls	Grady Hendrix
TX0008809460	My Heart Struck Sorrow	John Hornor Jacobs
TX0008114163	The Conformity	John Hornor Jacobs
TX0008830088	The Sea Dreams It Is The Sky	John Hornor Jacobs
TX0007905434	The Shibboleth	John Hornor Jacobs
TX0007561564	This Dark Earth	John Horner Jacobs
TX0002595154	A Pride of Princes	Jennifer Roberson
TX0003445261	A Tapestry of Lions	Jennifer Roberson
TX0005388531	Children of the Lion	Jennifer Roberson

Exhibit A

Registration Number	Title of Work	Author of Work
TX0002655076	Daughter of the Lion	Jennifer Roberson
TX0006608590	Deepwood: Karavans 2	Jennifer Roberson
TX0003038048	Flight of the Raven	Jennifer Roberson
TX0004575188	Highwaymen: Robbers & Rogues	Jennifer Roberson
TX0006494921	Karavans	Jennifer Roberson
TX0005700486	Lady of Sherwood	Jennifer Roberson
TX0003467465	Lady of the Forest (A Novel of Sherwood)	Jennifer Roberson
TX0004292057	Lady of the Glen	Jennifer Roberson
TX0002124929	Legacy of the Sword	Jennifer Roberson
TX0005388529	Legacy of the Wolf	Jennifer Roberson
TX0008936965	Life and Limb	Jennifer Roberson
TX0004840105	Mad Jack	Jennifer Roberson
TX0005450939 and TX0005450940	Out of Avalon	Jennifer Roberson
TX0006141041	Piece of Mind	Jennifer Roberson
TX0004248401	Return to Avalon: a celebration of Marion Zimmer Bradley	Jennifer Roberson
TX0002040293	Royal Captive	Jennifer Roberson
TX0001405603	Shapechangers	Jennifer Roberson
TX0005348007	Shapechanger’s Song	Jennifer Roberson
TX0009100077	Sinners and Saints	Jennifer Roberson
TX0001570167	Smoketree	Jennifer Roberson
TX0009259320	Sword-Bearer	Jennifer Roberson
TX0004752101	Sword-Born	Jennifer Roberson
TX0007724532	Sword-Bound	Jennifer Roberson

Exhibit A

Registration Number	Title of Work	Author of Work
TX0003187269	Sword-Breaker	Jennifer Roberson
TX0002125935	Sword-Dancer	Jennifer Roberson
TX0003031154	Sword-Maker	Jennifer Roberson
TX0002585643	Sword-Singer	Jennifer Roberson
TX0005504873	Sword-Sworn	Jennifer Roberson
TX0002496298	The Court of the Summer King	Jennifer Roberson
TX0005556700	The Lion Throne	Jennifer Roberson
TX0001840915	The Song of Homana	Jennifer Roberson
TX0007614823	The Wild Road	Jennifer Roberson
TX0002138857	Track of the White Wolf	Jennifer Roberson