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

Attorneys for Plaintiffs

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

SUSANA MARTINEZ-CONDE, STEPHEN L. MACKNIK, Plaintiffs, v. **APPLE INC.,** Defendant.	Case No. **CLASS ACTION COMPLAINT** **CLASS ACTION** **DEMAND FOR JURY TRIAL**

CONTENTS

Plaintiffs Susana Martinez-Conde and Stephen L. Macknik ("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 infringed upon Plaintiffs' and Class members' copyrights by reproducing their registered works without authorization as a part of amassing centralized databases of training materials and using that data to train its "Apple Intelligence" AI models on Plaintiffs' and Class members' copyrighted books and other works. Apple has created a set of generative AI models collectively called Apple Intelligence that it provides to consumers in its phones, tablets and personal computers. It used Plaintiffs' and Class members' copyrighted works without their authorization, without compensating them to train and test their Apple Intelligence models, a impermissible use far beyond any applicable license Apple has to sell such books to the users of its products.

2. Apple Intelligence is comprised of multiple generative AI models. These include but are not limited to "Apple Foundation Models," a ~3 billion parameter on-device language model, and a larger server-based language model, and Apple's OpenELM models, a family of "Open Efficient Language Models."

3. Apple reproduced and used data sets that included Books3, a dataset of pirated, copyrighted books that includes the published works of Plaintiffs and the Class, as training data for its AI models. Apple used the books in Books3, among others, to train its OpenELM language models and its Foundation Language Models.

4. Books3 is a notorious "shadow library," a dataset of pirated, copyrighted books that can be found in various places on the internet or shared and downloaded from pirate websites and file sharing protocols like BitTorrent, the popular peer-to-peer (P2P) file sharing protocol for copying and distributing infringing material. It is often found as a part of other datasets of pirated books.

5. Along with knowingly using training datasets that include Books3 to train its models, Apple uses "Applebot," a web-crawling software program that copies mass quantities of

internet data (also known as "scraping") to use as training data. Apple scraped data with Applebot for nearly nine years before disclosing that it intended to use the scraped data to train its AI systems. Web crawlers like Applebot scrape shadow libraries including but not limited to Books3, that host millions of other unlicensed copyrighted books, including Plaintiffs' and Class members' copyrighted works.

6. On information and belief, Apple also trained its models on unauthorized copies of eBooks it sells to its users through Apple Books. Copying and using such eBook files for any purpose beyond the explicit, limited scope of Apple's license to sell them is copyright infringement.

7. Generative AI models like those used in Apple Intelligence are only as good as the training data on which they are trained. Bad writing in training data results in less valuable, less useful AI models. Good writing in training data makes AI outputs better and models more valuable. This is why Apple and other AI companies prioritize and use high quality writing, like copyrighted works, to train and fine-tune their models.

8. Apple's AI models were created and operate by first making unauthorized digital copies of copyrighted works as a part of amassing central databases of enormous amounts of textual works and images to use for various purposes, including as training data. Apple then prepares such data for training AI models, which involves the creation of unauthorized copies and the preparation of derivative works in training data sets. Apple trains a model by using computers to analyze and record massive amounts of information about the relationships between words or bits of words ("tokens") in those works; using advanced mathematics and computer processing to predict a sequence of words in response to a text prompt, based on that detailed information about the relationship between tokens—the very creative expression found in those copyrighted works; and fine-tuning and mid-training the model on the most desired texts with the best writing, to achieve preferred model outcomes and outputs.

9. Plaintiffs and the Class are authors who have registered copyrights for their published works. They did not consent to the unauthorized reproduction by Apple and use of their works to be stored in databases for general use by Apple or for use in any Apple Intelligence

model, including the Foundation Intelligence Models and OpenELM language models. Such pilfered intellectual property, along with being used for AI training and fine-tuning, is used for testing model performance, and for the creation of filters to prevent model outputs containing recited or regurgitated copyrighted materials from reaching the end user.

10. Plaintiffs and the Class did not consent to Apple making further reproductions of their works or preparing altered derivative work based on their copyright works to use as training data for Apple intelligence models. Apple prepared derivative works by processing and modifying the raw training data they copied, including Plaintiffs' books, to create derivative training data sets used in training their Apple Intelligence models.

11. The market for licensing AI training data is growing rapidly. Licensing deals to use copyrighted works as training data between AI developers and publishers are regularly in the news. Nevertheless, Apple did not compensate creators for use of their copyrighted works and concealed the sources of their training datasets to evade legal scrutiny. Apple continues to retain private AI training-data, including pirated books, to train its future models in various datasets without seeking Plaintiffs' or Class members' consent or providing them compensation.

12. Apple has illegally copied Plaintiffs copyrighted works to train its AI models, whose outputs compete with and dilute the market for those very works—works without which Apple Intelligence would have far less commercial value. This conduct has damaged Plaintiffs' and Class members' intellectual property. It deprived Plaintiffs and the Class of control over their work, undermined the economic value of their copyrighted works, and positioned Apple to achieve massive commercial success through unlawful means.

II. JURISDICTION AND VENUE

13. This Court has subject-matter jurisdiction under 28 U.S.C. §§ 1331, 1332(d), 1400(a) because this case arises under the Copyright Act (17 U.S.C. § 501).

14. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. §§ 1391(c)(2) and 1400(a) because Apple is headquartered in this district.

III. DIVISIONAL ASSIGNMENT

15. Pursuant to Civil Local Rules 3-2(c) and 3-2(e), assignment of this case to the San Jose Division is proper because Apple is located in Santa Clara County where a substantial part of the events giving rise to Plaintiffs' and Class members' claims occurred.

IV. PARTIES

A. Plaintiffs

16. **Plaintiff Susana Martinez-Conde** is a Professor of Ophthalmology, Neurology, and Physiology & Pharmacology at SUNY Downstate Health Sciences University. Professor Martinez-Conde received a BSc in Experimental Psychology from Universidad Complutense de Madrid and a Ph.D in Medicine and Surgery from the Universidade de Santiago de Compostela in Spain. She was a postdoctoral fellow with the Nobel Laureate Prof. David Hubel, and then an Instructor in Neurobiology, at Harvard Medical School. Professor Martinez-Conde writes frequently for *Scientific American* and previously had a regular column in *Scientific American: MIND* on the neuroscience of illusion. She is the 2014 recipient of the Science Educator Award, a prestigious prize given by the Society for Neuroscience (30,000 members) to an outstanding neuroscientist who has made significant contributions to educating the public.

17. Prof. Martinez-Conde′s research has been featured in print in The New York Times, The New Yorker, The Wall Street Journal, Wired, The LA Chronicle, The Times (London), The Chicago Tribune, The Boston Globe, Der Spiegel, etc., and in radio and TV shows, including Discovery Channel′s Head Games and Daily Planet shows, NOVA: scienceNow, CBS Sunday Morning, NPR′s Science Friday, and PRI′s The World. She works with international science museums, foundations and nonprofit organizations to promote neuroscience education and communication.

18. **Plaintiff Stephen L. Macknik** is a Professor of Ophthalmology, Neurology, and Physiology & Pharmacology at SUNY Downstate Health Sciences University. Professor Macknik is Professor Martinez-Conde's writing partner and husband. Together, they authored the international bestselling book Sleights of Mind: What the Neuroscience of Magic Reveals About Our Everyday Deceptions, which has been published in 19 languages, distributed worldwide,

listed as one of the 36 Best Books of 2011 by The Evening Standard, London, and received the Prisma Prize to the best science book of the year. Sleights of Mind appears in Books3 and Apple created unauthorized copies of it and used its to train its AI models that compete with Plaintiffs and diminish the value of their works and their labor.

19. Professor Macknik completed a triple-major in Psychobiology, Biology, and Psychology at the University of California, Santa Cruz in 1991. Thereafter, he completed his PhD in Neurobiology at Harvard University in 1996. He received his postdoctoral training from the Nobel Laureate Prof. David Hubel at Harvard Medical School, from 1996 to 2001. His research and writing have been featured in print in *The New York Times*, *The New Yorker*, *The Wall Street Journal*, *The Atlantic*, *Wired*, *The LA Chronicle*, *The Times (London)*, *The Chicago Tribune*, *The Boston Globe*, *Der Spiegel*, and in radio and TV shows, including *Discovery Channel's Head Games* and *Daily Planet* shows, *NOVA: scienceNow*, *CBS Sunday Morning*, *NPR's Science Friday*, and *PRI's The World*.

20. Professors Martinez-Conde and Macknik have multiple copyright registrations including:

Full Title	**Registration No.**	**Date**
CHAMPIONS OF ILLUSION: The Science Behind Mind-Boggling Images and Mystifying Brain Puzzles	TX0008595717	2/2/2018
SLEIGHTS OF MIND: What the Neuroscience of Magic Reveals About Our Everyday Deceptions	TX0007300820	12/27/2010

21. Sleights of Mind is included in the Books3 dataset. Apple copied Sleights of Mind without authorization, and used it to train its AI models. Champions of Illusion can be found in Library Genisis shadow library.

B. Defendant

22. 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

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

24. 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 Defendants with respect to the acts, violations, and common course of conduct alleged herein.

V. FACTUAL ALLEGATIONS

25. Apple is one of the three largest companies in the world with a $3.8 trillion market capitalization. Apple is an electronics and media company that designs, manufactures, and sells software and hardware products, including the ubiquitous iPhone. Every second, Apple sells seven iPhones. It sells television, music, movies, podcasts, books and other media to consumers through its proprietary applications on its phones, tablets and computers. Indeed, Apple has licenses related to many of the digital book versions of the books at issue in this lawsuit, for the limited purpose of selling them to consumers through their Apple Books app. Any use of such books beyond that purpose is infringement.

26. Apple is also building commercial AI products. 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."

27. In or around June 2024, Apple announced the development of its commercial artificial intelligence platform, called Apple Intelligence.[1] Apple Intelligence includes multiple generative-AI models and related tools and technologies. The day after Apple officially introduced Apple Intelligence the company gained more than $200 billion in value: "the single most lucrative day in the history of the company."

28. To train the generative-AI models that are part of Apple Intelligence, Apple first amassed an enormous library of raw training data. Apple scraped the internet with a web-crawler called "Applebot." It also utilized training datasets that contained pirated books in the Books3 shadow library. Part of Apple's data library includes unauthorized copies of Plaintiffs' and Class members' copyrighted works that were copied without authorization.

29. Apple has not attempted to pay these authors for their contributions to their commercial AI products that compete directly with them and their copyrighted works. Apple did not seek licenses to copy and use the copyrighted books it used to train to its models. Instead, it intentionally evaded payment by using books already compiled in pirated datasets.

A. How large language models work.

30. Artificial intelligence—commonly abbreviated "AI"—denotes software that is designed to algorithmically create an illusion of human reasoning or inference, often using statistical and mathematical methods.

31. The Apple Intelligence platform includes multiple AI software programs called large language models ("LLMs") that Apple created for use in its products. Apple created these generative models for use in a wide range of AI features integrated across its platforms. An LLM is AI software designed to emit convincingly naturalistic text outputs in response to user prompts.

32. Apple's LLM training started by amassing large quantities of raw, pre-training data. Apple admits that it sources a significant portion of the pre-training data for its models from web content crawled by Applebot, spanning hundreds of billions of links and pages, covering an extensive range of languages, locales, and topics. Apple represents that AppleBot "employs

[1] See https://machinelearning.apple.com/research/introducing-apple-foundation-models

advanced crawling strategies to prioritize high-quality and diverse content," and that "high-quality filtering plays a critical role in overall model performance." Apple's web crawler made unauthorized reproductions of online pirated book libraries known as "Shadow Libraries" including the "Books3" shadow library in which Plaintiffs' works are included.

33. Apple also curates and uses training datasets that include copied shadow libraries, including Books3, to train its models.

34. Apple then processes that pre-training data for use as model training data. Raw datasets are processed and filtered in various ways in pretraining, creating derivative datasets that are fed into models for training analysis. Apple applies filters to remove certain categories of personally identifying information, copyright notices and license language, unwanted language, profanity and unsafe material. This corpus of text is called the "training dataset."

35. An LLM is "trained" by analyzing text data to extract massive amounts of information about the relationships between each "token" in a textual work. A token is a small string of characters, a word or sometimes a piece of a word, that serves as the base unit of AI models. AI models process the relationships between each token in a work and stores that information in what are called "weights." The LLM copies and analyzes each textual work in the training dataset and extracts the protected expression from it in the finest detail, on a token-by-token basis. The LLM records the results of this process for each token 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.

36. Weights are the fundamental control knobs or settings for a model and determine how a model evaluates new data and makes predictions. They are the core parameters for an LLM and are learned during training. LLMs can have millions or even billions of weights. Weights are numerical variables that set the relative importance of the features in the dataset on the output. They are based entirely on the relationships between the words of copyrighted materials.

37. 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 by tokenizing that prompt and using mathematics to generate the most probable string of tokens in response to that prompt. Whenever an LLM generates a response to a user prompt, it is performing a computation that relies on these stored weights recorded from copies of Plaintiffs' and Class members' works, and is imitating the protected expression ingested from the training dataset. There is no human ingenuity or creativity in any model output. It is a computer program performing mathematical operations using information it copied and extracted from its training data.

38. LLMs have a universal problem in "memorizing" their training data and regurgitating or reciting their training data in outputs from time to time. Apple Intelligence models are no different. During training, an LLM processes pieces of text in its training data and learns probabilistic relationships across the dataset as a whole. These learned patterns are not so high-level or abstract; rather, they are often highly specific. Training data is not so transformed by training, rather it is sometimes "memorized" by the model—encoded in some form inside the LLM's weights. Such memorized content can sometimes be "extracted" later; they can be reproduced in an LLM's outputs at generation time.[2]

39. The ability to extract verbatim portions of training data generates a copy of training data, but it also demonstrates the existence of a copy of that training data is memorized inside the model itself. The model itself also being a copy has important implications. Notably, a model—not just extracted training data—is an infringing copy of the training data it has memorized. At minimum, the copied data Apple integrates to prevent regurgitation or recitation of training data in Apple Intelligence models is an unauthorized copy of Plaintiffs' and Class members' works. Copyright law offers the destruction of infringing materials as a remedy.

B. Apple trained its OpenELM models on copyrighted works.

40. 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.

[2] See https://arxiv.org/pdf/2404.12590

OpenELM uses a layer-wise scaling strategy to efficiently allocate parameters within each layer of the transformer model, leading to enhanced accuracy."

41. 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 tokens and 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") parameters (the weights and biases) recorded from the protected expression found in its training dataset.

42. 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."[3]

43. Apple's GitHub repository confirms "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."[4]

44. The Pile (Gao et al., 2020; Biderman et al., 2022), is a curated collection of English language datasets including Books3 used that is popular for training large LLMs. It was curated by a research organization called EleutherAI. Books3 is a component of The Pile.[5]

[3] See https://huggingface.co/apple/OpenELM

[4] See https://github.com/apple/corenet/blob/main/projects/openelm/README-pretraining.md

[5] See https://arxiv.org/pdf/2201.07311v1

45. 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:[6]

> 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.

46. Apple also published a paper about OpenELM ("OpenELM Paper").[7] In a table called "Dataset used for pre-training OpenELM," Apple reveals that a large quantity of training

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.

data comes from the "Books" subset of a dataset called "RedPajama."

47. Apple's OpenELM models were trained on RedPajama-V1.[8] RedPajama-V1 "is a publicly available, fully open, best-effort reproduction of the training data…used to train the first iteration of LLaMA family of models." This LLaMA training dataset included Books3 section of The Pile.[9]

[6] See https://arxiv.org/pdf/2101.00027

[7] See https://arxiv.org/pdf/2404.14619

[8] See https://arxiv.org/html/2411.12372v1.

[9] See https://arxiv.org/pdf/2302.13971.

48. 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.

49. Bibliotik is one of several notorious shadow libraries, along with 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.

50. The person who assembled the Books3 dataset, Shawn Presser, has confirmed in public statements that it represents "all of Bibliotik" and contains approximately 196,640 books.

51. 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.

52. By using the entire text, Apple made and used copies of each book in Books3 to store as reference data and to train its OpenELM models and develop other analytic and filtering tools.

53. Plaintiffs' copyrighted works are among the works in the Books3 dataset.

54. 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."

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

56. 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 Apples's own information revealed in the OpenELM research paper and on its model card on Hugging Face, this includes Apple.

57. Apple has admitted training its OpenELM large language models on a copy of the "Books" subset of the RedPajama dataset, as well as a deduplicated version of The Pile, each of which in turn contains a copy of the Books3 dataset. Therefore, Apple trained its OpenELM models on a copy of Books3, a known body of pirated books.

58. Because Plaintiffs' copyrighted book is part of Books3, Apple copied in its entirety without authorization, and trained OpenELM, on one or more copies of the Plaintiffs copyrighted works and directly infringed Plaintiffs' copyrights along with the copyrights of the Class.

C. Apple trained its Foundation Language Models on copyrighted works.

59. Apple announced their two Apple Intelligence Foundation Language Models in June 2024. These models were described in a paper of the same name released by Apple on July 29, 2024 (the "FLM Paper").[10]

60. 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."

61. The FLM Paper emphasizes the special importance of a foundation model's capacity to write: "Writing is one of the most critical abilities for large language models to have, as it empowers various downstream use cases such as changing-of-tone, rewriting, and summarization."

62. 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

[10] See https://arxiv.org/pdf/2407.21075

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.)". This means that both the AFM-server and AFM-on-device models are trained on the same corpus of training data.

63. 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."

64. 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."

65. 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 we filtered the datasets to remove personally identifiable information, copyright management information, and other undesired text before including them in the pre-training mixture.

66. 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. In practice, "publicly available" means works that can be downloaded somewhere from the public internet, which contains a vast number of copyrighted works by authors who have not granted a license for reproduction. There is a name for this kind of copying: copyright infringement. There is also a name for the infringing copies: pirated works.

67. For instance, 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 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.

68. 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 someone other than the author made curated copies of copyrighted works freely available, without the author's permission. These are just pirated works included in a curated dataset that the pirate claims is open-source.

69. 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' consent. Only the copyright owner can offer their copyrighted work to the public under an open-source license. A third party cannot usurp that right.

70. Therefore, when Apple says that a major source of the training data for its Foundation Language Models is "publicly-available or open-sourced datasets," they are talking about curated datasets like RedPajama. 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, Apple's reference to "publicly-available or open-sourced datasets" likely includes Books3, and that Apple therefore included Books3 in the training dataset for its Foundation Language Models.

71. Plaintiffs' and Class members' copyrighted works are part of Books3. It follows that Apple trained its Foundation Language Models on one or more copies of each book therin, thereby directly infringing the copyrights of the Plaintiffs and the Class. Apple has created a permanent AI training data library containing copies of all these "publicly-available or open-sourced datasets" in expectation of training future models.

72. As to its 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." Applebot has been crawling the web since approximately mid-2015. 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.

73. 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.

74. 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 scraped web pages. These model-based classifiers are themselves trained on datasets that include unlicensed copyrighted works.

75. 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,"[11] 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, the part of Apple's training dataset that comes from filtered Applebot pages includes copyrighted works from commercial news and media websites.

[11] See https://papers.ssrn.com/sol3/Delivery.cfm/5009668.pdf

76. 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 directly.

77. Apple obscures the training datasets for its Apple Intelligence Foundational Language Models to blur its use of copyrighted materials. Apple's decision not to disclose the training datasets for Apple Intelligence 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.

78. The "curated publicly available or open-sourced datasets" that Apple copied for the training datasets for Apple Intelligence contain copyrighted material, including Plaintiffs' copyrighted works. Such use of datasets with copyrighted works would be consistent with Apple's process for training its OpenELM model. Apple described its training data for OpenELM, including data from The Pile, as "public datasets." 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.

79. There are numerous examples of publicly reported AI licensing deals. 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 data set licensing companies have formed a trade group called the Dataset Providers Alliance. 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.

80. 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.

81. 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 trainings its models.

82. Furthermore, Apple is reportedly exploring a paid tier for users of its Apple Intelligence products. Doing so might be in effort to offset the costs of its steep investments in building Apple Intelligence. Analysts contend that Apple Intelligence could add $4 trillion to the company's market capitalization.

D. Apple's conduct impairs the market for Plaintiffs' and Class members' works.

83. Apple has neither paid nor sought permission from Plaintiffs for the use of their copyrighted works. Instead, Apple downloaded, scraped, or otherwise copied vast quantities of copyrighted works—including illegally compiled datasets such as Books3—that included Plaintiffs' works. Apple has diminished the value of Plaintiffs' intellectual property by making unauthorized copies and derivative works for use in training their Apple Intelligence models. Apple further deprived Plaintiffs of the revenue that would have been generated had Apple approached Plaintiff or their licensing agents directly to license copies of their works for use in AI training. Furthermore, compiling private libraries sourced from illegally compiled datasets for AI training purposes may "lead to a loss of sales" by "harm[ing] the market for access to those works."

84. Apple's unauthorized use of Plaintiffs' copyrighted works creates a risk of market dilution of the works themselves and of Plaintiff's professional writing. Works generated using Apple Intelligence will inevitably start competing with Plaintiffs' copyrighted and other works and ultimately dilute royalty pools and professional writing opportunities as AI-generated output increasingly floods the market. Already, "low-quality sham 'books'" 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.

85. Apple's unauthorized use of Plaintiffs' copyrighted works to train Apple Intelligence models has caused and threatens to cause substantial harm to the actual potential markets for those works. Plaintiffs and similarly situated creators previously licensed their work for their own commercial uses. Apple's conduct has disrupted this traditional market and impaired the emergence of lawful licensing regimes by obtaining and exploiting authors' works without consent or compensation.

86. Apple's models generate outputs by automated computer operation that substitute for 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 and reduced recognition, and sales among other harms.

VI. CLASS ACTION ALLEGATIONS

87. The "Class Period" as defined in this Complaint begins at least three years before the date of this complaint's filing and runs through the present.

88. Apple engaged in a course of conduct common to all class members in infringing Plaintiffs' and Class members' works including:

a. Apple reproduced all class member books in the Books3 dataset without authorization to train the OpenELM models;

b. Apple further reproduced all class member books in the Books3 dataset without authorization to train its Foundation Language Models;

c. Apple made unauthorized copies and prepared of unauthorized derivative works of all class member books in the course of pre-training for those models;

d. Apple made and used unauthorized copies of datasets that included class members' unlicensed copyrighted works to train classifier modes; and

e. Apple retained copies of all the training data it has gathered and processed thus far without authorized, in the form of a private data library for potential use in future models—an AI training data library —that includes the Books3 dataset, which, in turn, includes the Plaintiffs' and Class members' infringed works.

89. 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 legal or beneficial owners of a registered copyright for any work Apple copied without authorization or ingested for training or otherwise developing its OpenELM Models and Foundation Models during the relevant time period, which was registered with the United States Copyright Office within five years of the work's publication and which was registered with the United States Copyright Office before being trained on by Apple, or within three months of publication. Excluded from the Classes are the Defendant, its subsidiaries and affiliates, officers, executives, and employees; Defendant's attorneys in this case, federal government entities and instrumentalities, states or their subdivisions, and all judges and jurors assigned to this case.

90. Plaintiffs reserve the right to modify or amend the definition of the proposed Class before the Court determines whether certification is appropriate.

91. The Class members are so numerous and geographically dispersed that individual joinder of all Class members is impracticable. Moreover, given the costs of complex antitrust litigation, it would be uneconomical for many Plaintiffs to join their individual claims. 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.

92. 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, or other datasets used to train one or more of the Apple Intelligence models.

93. The Class can further be identified by analyzing the training data that Apple used for both its OpenELM Models and Apple Foundation Models.

94. ISBN numbers can ne used to identify copyrighted books in Apple's training data.

95. 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.

96. Plaintiffs will fairly and adequately protect and represent the interests of the Class. The interests of the Plaintiffs are aligned with, and not antagonistic to, those of the other Class members. Further, Plaintiffs have retained competent counsel who are experienced in litigating federal class actions and other complex class action litigation involving copyright infringement in the context of training AI models.

97. Numerous questions of law and fact are common to each Class member arising from Defendant's conduct, including:

a. Whether Plaintiffs' and Class members' works were included in the training datasets used by Apple to train its Apple Intelligence product, including the RedPajama and Pile datasets Defendant used;

b. Whether Apple's inclusion of Plaintiffs' and Class members' works in their training datasets constituted or required the works' unauthorized reproduction by Apple;

c. Whether Apple lacked authorization to reproduce or make copies of Plaintiff's and Class members' works;

d. Whether Apple violated Plaintiff's and the Class members' copyrights when it downloaded copies of Plaintiff's copyrighted works and used them in training its OpenELM and Foundation Models;

e. Whether Apple violated Plaintiff's' and Class members' copyrights when it downloaded copies of Plaintiff's copyrighted works without authorization and used them in its Apple Intelligence products;

f. Whether Apple violated Plaintiff's and Class members' copyrights by creating unauthorized copies or derivative works based on their copyrighted works in pre-training and training of the OpenELM and Foundation Models;

g. Whether Apple violated Plaintiff's and Class members' copyrights by creating unauthorized copies or derivative works based on their copyrighted works in creating the OpenELM and Foundation Models themselves, or the filters they employ to prevent the regurgitation of copyrighted works in model outputs;

h. Whether Apple violated Plaintiff's and Class members' copyrights by creating

unauthorized copies or derivative works based on their copyrighted works in pre-training and training of the OpenELM and Foundation Models;

i. Whether this Court should enjoin Defendant from engaging in the unlawful conduct alleged herein or provide other equitable relief;

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

k. Whether Apple's infringement was willful; and

l. the appropriate measure of damages.

98. These and other questions of law and fact are common to the Class and predominate over questions affecting Class members on an individual basis. Damages pose no obstacle for class certification as statutory damages are available to all Class members pursuant to 17 U.S.C. § 504.

99. 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. The alternative of separate actions by individual Class members risks inconsistent adjudications and is an inefficient use of limited judicial resources.

100. Plaintiffs know of no difficulty to be encountered in the maintenance of this action that would preclude its maintenance as a class action.

VII. CLAIMS

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

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

102. As the owners of the registered copyrights in their copyrighted books and other copyrighted works, Plaintiffs and Class members hold the exclusive copyrights in those works under 17 U.S.C. § 106, including the exclusive right to: (1) reproduce the copyrighted work in

copies; (2) prepare derivative works based upon the copyrighted work; (3) distribute copies of the copyrighted work; (4) perform the copyrighted work; and (5) display the copyrighted work to the public. Plaintiffs and the Class members never authorized Apple to make copies of their copyrighted books and other copyrighted works, prepare 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.

103. Apple made all its copies of Plaintiffs' copyrighted books and other copyrighted works without Plaintiff's or Class members' permission, violating their exclusive rights under the U.S. Copyright Act. Indeed, "the person who copies the textbook from a pirate website has infringed already, full stop." *Bartz et al. v. Anthropic*, No. C 24-05417 WHA, 2025 WL 1741691 at *11 (N.D. Cal. June 23, 2025). Regardless of how Apple uses the works in its private training-data library in the future, this cannot negate the initial copying of works sourced from shadow libraries infringed on Plaintiff's and Class members' exclusive rights.

104. Plaintiffs and Class members have been injured by the Apple's direct copyright infringement of the their copyrighted works. Plaintiffs and Class members are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law or in equity.

VIII. PRAYER FOR RELIEF

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 Plaintiff's 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 Intelligence or other LLM models and training sets that incorporate Plaintiff's and Class members' works;

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

f. 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: October 9, 2025

Respectfully Submitted,
JOSEPH SAVERI LAW FIRM, LLP.

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

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

Attorneys for Plaintiffs