Benjamin Gould (SBN 250630)
KELLER ROHRBACK L.L.P.
1201 Third Avenue, Suite 3400
Seattle, WA 98101-3268
(206) 623-1900, Fax (206) 623-3384
bgould@kellerrohrback.com

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

Additional Counsel Listed in Signature Block

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

GRADY HENDRIX & JENNIFER ROBERSON, *Individual and Representative Plaintiffs*, v. APPLE INC., *Defendant*.	No. **CLASS ACTION COMPLAINT** **JURY TRIAL DEMANDED**

Plaintiffs Grady Hendrix and Jennifer Roberson (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 Intelligence is a set of generative AI programs and technologies designed and maintained by Apple.

2. Apple—one of the world's most valuable companies—has invested substantial capital and engineering resources into Apple Intelligence. It 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.

3. But Apple is building part of this new enterprise using Books3, a dataset of pirated copyrighted books that includes the published works of Plaintiffs and the Class. Apple used Books3 to train its OpenELM language models. Apple also likely trained its Foundation Language Models using this same pirated dataset.

4. Apple is building another part of its Apple Intelligence empire by using 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 AI systems on this scraped data. Scrapers like Applebot can also reach "shadow libraries" that host millions of other unlicensed copyrighted books, including, on information and belief, Plaintiffs' and Class Members' copyrighted works.

5. The Foundation Language Models within Apple Intelligence depend on the contents of their training datasets. The Foundation Language Models 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 authors who have registered copyrights for their published works. They did not consent to the use of their works in any Apple Intelligence model, including the Foundation Intelligence Models and OpenELM language models.

7. The licensing market for AI training data is burgeoning. Nevertheless, Apple did not compensate creators for use of their copyrighted works and concealed the sources of their training datasets to evade legal scrutiny. On information and belief, Apple continues to retain a private AI training-data library including thousands of pirated books to train its future models, without seeking Plaintiffs' or Class Members' consent or providing them compensation.

8. In sum, 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—works without which Apple Intelligence 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.

II. JURISDICTION AND VENUE

9. 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).

10. 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

11. Under Civil Local Rule 3-2(c), assignment of this case to the San Francisco 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

12. Plaintiff Grady Hendrix is an author who lives in New York. He owns registered copyrights in multiple books, including *My Best Friend's Exorcism*, U. S. Copyright Registration TX0008276604) (the "Hendrix Asserted Work").

13. Plaintiff Jennifer Roberson is an author who lives in Arizona. She owns registered copyrights in multiple books, including *Sword-Bound*, U. S. Copyright Registration TX0007724532 (the "Roberson Asserted Work").

14. Below, the Hendrix Asserted Work and Roberson Asserted Work are referred to collectively as the Infringed Works. On information and belief, Apple may have also infringed other works by Plaintiffs.

B. Defendant

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

V. FACTUAL ALLEGATIONS

16. Apple, one of the biggest companies 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. 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."

17. 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."

18. To train the generative-AI 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 consent, credit, or compensation.

19. Apple has not attempted to pay these authors for their contributions to this potentially lucrative venture. Apple did not seek licenses to copy and use the copyrighted books provided to its models. Instead, it intentionally evaded payment by using books already compiled in pirated datasets.

A. How large language models work.

20. 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.

21. The Apple Intelligence platform includes multiple AI software programs called large language models ("LLMs") that have been created, maintained, and commercialized by Apple. An LLM is AI software designed to emit convincingly naturalistic text outputs in response to user prompts.

22. Though an LLM is a software program, it is not created the way most software programs are—by human software programmers writing code. Rather, an LLM is trained by copying an enormous quantity of textual works and then feeding these copies into the model. This corpus of text is called the "training dataset."

23. 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.

24. 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 and imitating the protected expression ingested from the training dataset.

25. Much of the material in Apple's training dataset, however, 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.

26. 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."

27. 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.

28. 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 … a subset of RedPajama."

29. 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.

30. 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.

31. Once again, though, 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.

32. Bibliotik is one of several notorious "shadow library" websites that also includes 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.

33. 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.

34. 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.

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

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

37. 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."

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

39. 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, this includes Apple.

40. 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.

41. 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.

42. In June 2024, Apple announced its commercial AI software platform, called Apple Intelligence. Apple Intelligence includes several AI models developed by Apple. Two of these models are called the *Apple Intelligence Foundation Language Models*. These models were described in a paper of the same name released by Apple on July 29, 2024 (the "FLM Paper").

43. 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."

44. 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]."

45. 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.

46. 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."

47. 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."

48. 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."

49. 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.

50. 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.

51. 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.

52. 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 holder can offer their copyrighted work to the public under an open-source license. A third party cannot usurp that right.

53. 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 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, it is, on information and belief, likely that Apple's reference to "publicly-available or open-sourced datasets" includes Books3, and that Apple therefore included Books3 in the training dataset for its Foundation Language Models.

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

55. 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.

56. 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.

57. 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. On information and belief, these model-based classifiers are themselves trained on datasets that include unlicensed copyrighted works.

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

59. 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.

60. On information and belief, Apple obscures the training datasets for its Apple Intelligence Foundational Language Models to blur its use of copyrighted materials. On information and belief, 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. Apple has neither paid nor sought permission from Plaintiffs for the use of their copyrighted works. Instead of doing so, Apple downloaded, scraped, or otherwise copied vast quantities of copyrighted works—including illegally compiled datasets such as Books3—that included Plaintiffs' works like the Infringed Works. Apple has 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. 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."

67. Apple's unauthorized use of the Infringed Works creates a risk of market dilution. Works generated using Apple Intelligence 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'" 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.

68. 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.

69. Apple's models 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 and reduced recognition, and sales among other harms.

VI. CLASS ACTION ALLEGATIONS

70. 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.

71. As used here, the term "Apple Intelligence Tasks" refers collectively to Apple's numerous separate infringing uses of Plaintiffs' and Class Members' works including the Infringed Works:

A. Apple's unauthorized reproduction of the Books3 dataset—which includes the Infringed Works—to train its OpenELM language models;

B. Apple's further unauthorized reproduction of the Books3 dataset, which includes the Infringed Works, to train its Foundation Language Models;

C. Apple's unauthorized use of datasets that included unlicensed copyrighted works to train classifier model; and

D. Apple's unauthorized retention of all the training data it has gathered and processed thus far, 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 Infringed Works.

72. **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 persons or entities domiciled in the United States that own a registered United States copyright in any work that was used for one or more of the Apple Intelligence Tasks during the Class Period.**

73. 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.

74. **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.

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

76. The Class can further be identified by analyzing the training data that Apple used for both its OpenELM model and Apple Intelligence Foundational Language Model.

77. **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.

78. **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.

79. **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 Plaintiffs' and Class Members' works were included in the training datasets used by Apple to train its Apple Intelligence product, including the RedPajama datasets Defendant used;

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

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

D. Whether Apple violated the copyrights of Plaintiffs and the Class when it downloaded copies of Plaintiffs' Infringed Works and other copyrighted works and used them in training its OpenELM model;

E. Whether Apple violated the copyrights of Plaintiffs and the Class when it downloaded copies of Plaintiffs' Infringed Works and other copyrighted works and used them in its Apple Intelligence products;

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

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

H. Whether Apple's infringement was willful.

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

81. **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.

82. **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

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

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

84. 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.

85. In conducting the Apple Intelligence Tasks, 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 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 that the initial copying of works sourced from shadow libraries infringed on Plaintiffs' and Class Members' exclusive rights.

86. Plaintiffs and Class Members have been injured by the Apple Intelligence Tasks, each of which constitutes direct copyright infringement of the Infringed Works. Plaintiffs and Class Members are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

VIII. PRAYER FOR RELIEF

87. 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 Intelligence or other LLM models and training sets that incorporate Plaintiffs' 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 this 5th day of September, 2025.

KELLER ROHRBACK L.L.P.

By *s/ Benjamin Gould*

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

By *s/ Matthew Butterick*

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

Attorneys for Plaintiffs