JOSEPH W. COTCHETT (SBN 36324)
BRIAN DANITZ (SBN 247403)
GIA JUNG (SBN 340160)
CAROLINE YUEN (SBN 354388)
COTCHETT, PITRE & MCCARTHY, LLP
840 Malcom Road
Burlingame, CA 94010
Tel: 650-697-6000
jcotchett@cpmlegal.com
bdanitz@cpmlegal.com
gjung@cpmlegal.com
cyuen@cpmlegal.com

KARIN B. SWOPE (*pro hac vice* forthcoming)
THOMAS LOESER (SBN 202724)
ANDREW FULLER (*pro hac vice* forthcoming)
JACOB ALHADEFF (*pro hac vice* forthcoming)
COTCHETT, PITRE & MCCARTHY, LLP
1809 7th Avenue, Suite 1600
Seattle WA 98101
Tel: (206) 802-1272
kswope@cpmlegal.com
tloeser@cpmlegal.com
afuller@cpmlegal.com
jalhadeff@cpmlegal.com

Counsel for Plaintiff

LESLEY E. WEAVER (SBN 191305)
ANNE K. DAVIS (SBN 267090)
JOSHUA D. SAMRA (SBN 313050)
BLEICHMAR FONTI & AULD LLP
1330 Broadway, Suite 630
Oakland, CA 94612
Tel. (415) 445-4003
lweaver@bfalaw.com
adavis@bfalaw.com
jsamra@bfalaw.com

UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA
SAN JOSE DIVISION

TASHA ALEXANDER, an Individual on Behalf of Herself and All Others Similarly Situated,

Plaintiff,

v.

APPLE INC., a California Corporation, **CRAIG FEDERIGHI**, an individual, and **JOHN GIANNANDREA**, an individual,

Defendants.

Case No. 5:25-cv-09090

CLASS ACTION COMPLAINT

JURY TRIAL DEMANDED

TABLE OF CONTENTS

Plaintiff Tasha Alexander ("Plaintiff"), individually and on behalf of all others similarly situated, by and through their undersigned counsel, bring this class action complaint against Apple Inc. ("Apple"), Craig Federighi ("Federighi"), and John Giannandrea ("Giannandrea" and collectively, "Defendants").

I. INTRODUCTION

1. Apple, despite being one of the world's most profitable companies, willfully infringed millions of registered copyrighted works to build its commercial artificial intelligence ("AI") product, Apple Intelligence.

2. Apple Intelligence is a collection of writing, image, personal assistance, and other tools integrated into Apple products, powered by generative AI, and developed and maintained by Defendant Apple.

3. Plaintiff and Class Members, authors who have copyright registrations for their works, bring this action under the Copyright Act to address harm from Apple's willful infringement. Apple scrapes the internet and downloads pirated copies of their works, reproduces these works, and uses them to train its models to generate human-like text and images, which has and is displacing and diluting the market for Plaintiff's works.

4. Apple did not seek permission or pay for the copyrighted works it uses, rather Apple willfully exploits the works without authorization to build and train its Apple Intelligence product.

5. Authors invest years in creating these works, and intellectual property rights as old as our Constitution guarantee their right to compensation. Yet Apple disregards copyright protections, making large-scale infringement a core part of its business model for its flagship Apple Intelligence product.

6. Apple has committed significant capital and engineering expertise to developing Apple Intelligence and to support Apple Intelligence with infrastructure investments. Apple demonstrates its views that technology is a transformative innovation, by embedding Apple Intelligence in its products, aiming to fundamentally enhance user experiences across its product ecosystem. With Apple Intelligence, Apple aspires to avoid falling behind competitors that have

previously released generative AI models and to drive substantial growth, adding trillions to its market capitalization in the years ahead.

7. Despite the importance of Apple Intelligence to its future growth, Apple has made no effort to compensate Plaintiff and Class Members and has taken steps to conceal the extent of its copyright infringement. Copyright law prohibits Apple's actions—downloading and copying hundreds of thousands of copyrighted books from bootlegged pirated sources.

8. Apple undermines authors' livelihoods by embedding Apple Intelligence in its products, enabling Apple device users to generate, freely or cheaply, texts writers would otherwise be paid to produce. These models, which erode the market for Plaintiff's and the Class's works, were built without compensating the authors on whom Apple relied, without authorization, for high quality training data.

9. Apple has acknowledged that it used datasets such as RedPajama, C4, and PILE,[1] which are known to include shadow libraries of bootlegged copyrighted books such as Books3[2] to train its OpenELM language models.

10. Apple has also disclosed some of the data used to train its Foundation Language Models, which like the OpenELM models, use publicly available or open-sourced datasets.[3]

11. In addition to collecting datasets to train its generative AI models, Apple deployed Applebot, "the webcrawler for Apple,"[4] a software tool that systematically crawls websites and copies large volumes of web pages, including files available through webpages such as those posted to shadow libraries, a process known as "scraping."

---

[1] Sachin Mehta et al., *OpenELM: An Efficient Language Model Family with Open Training and Inference Framework*, MACHINE LEARNING RESEARCH AT APPLE (May 2, 2024), https://arxiv.org/pdf/2404.14619 (Apple's "OpenELM Paper").

[2] Red Pajama: https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1; https://www.together.ai/blog/redpajama-data-v2; The Pile: An 800GB Dataset of Diverse Text for Language Modeling https://arxiv.org/pdf/2101.00027; Books3: https://x.com/theshawwn/status/1320282149329784833?lang=en.

[3] Apple Intelligence Foundation Language Models, arXiv:2407.21075 [cs.AI], available at https://arxiv.org/pdf/2407.21075 (last visited Oct. 22, 2025).

[4] "About Applebot - Apple Support," available at https://support.apple.com/en-us/119829 (last visited Oct. 22, 2025).

12. Large Language Models (LLMs) take in user prompts and generate responsive content as output. LLM powered generative AI products such as Apple's require vast amounts of high quality data for training in order to provide high quality output in response to user prompts. The success of an LLM such as Apple's OpenELM and FLM depends heavily on the quality and contents of the datasets used for training, making copyrighted works an attractive source of high-quality, expressive content.

13. For example, to secure access to high quality content for its generative AI products, in 2024, Apple signed a multimillion-dollar licensing deal with Shutterstock. However, no such agreement was made with, or compensation was given to Plaintiff or the Class.

14. Apple did not offer website owners the option to opt-out of Applebot's scraping of website content and use by Apple to train its generative AI models until in or around the summer of 2024.

15. Plaintiff and the Class are artists and authors who hold copyright protecting their works from unauthorized use. Plaintiff and the Class never authorized the use of their works by Apple for use in training any of the models powering Apple Intelligence, including OpenELM and FLM models. Apple's unauthorized copying of the Plaintiff Works continues to harm Plaintiff through Apple's ongoing monetization of those works. Apple has integrated its Apple Intelligence products into an expanding set of Apple products, including iPhone, iPad, iPad mini, iMac, Mac; operating systems beginning with iOS 18, iPadOS 19, and macOS Sequoia and continuing to iOS 26, iPadOS 26, macOS Tahoe 26, watchOS 26, and visionOS 26; Apple Vision Pro and Apple Watch; and has embedded additional features over time, including Image Playground, Genmoji, Writing Tools, visual intelligence, Shortcuts, and Live Translation, powering each new Apple product or feature with infringing copies of Plaintiff's works.

16. Apple was able to develop Apple Intelligence because of the Generative AI Models it built and trained on Plaintiff's copyrighted works without Plaintiff's license or authorization.

17. By embedding Plaintiff's works into Apple Intelligence, Apple has irreversibly entangled Plaintiff's works with its commercial products, stripping Plaintiff of her exclusive rights under the Copyright Act to control the copying and distribution of the Plaintiff Works.

18. This unauthorized copying and distribution inflicts immediate and irreparable harm on Plaintiff—harm that Apple compounds daily through its expanding deployment of Apple Intelligence features embedded in Apple's products.

19. It is no secret that the datasets Apple relied on to train the models powering Apple Intelligence contained pirated and bootlegged copies of copyrighted works. Indeed, shadow library websites such as LibGen and Z-library were and are frequently seized by the FBI and international law enforcement partners, with prominent notices of seizure going up on the seized domains, replacing the searchable collections of pirated works, only for the shadow libraries to pop up on new domains.[5]

20. In short, Apple copied Plaintiff's and the Class's copyrighted works to create a massive commercially valuable dataset and to train AI models that compete with and displace those works—without which Apple Intelligence would be far less valuable. This has stripped

[5] Woodcock, Claire, *Shadow Libraries' Are Moving Their Pirated Books to The Dark Web After Fed Crackdowns*, VICE (Nov. 30, 2022), available at https://www.vice.com/en/article/shadow-libraries-are-moving-their-pirated-books-to-the-dark-web-after-fed-crackdowns/.

Plaintiff of her rights and harmed the economic value of their labor, enabling Apple's unlawful commercial gain.

II. PARTIES

A. Plaintiff

21. Plaintiff Tasha Alexander is an author who is the pseudonym of Anastasia Grant. She owns registered copyrights in multiple books which Defendant copied through direct downloading Books3 (the "Infringed Works").

Title	Registration No.	Date
And Only to Deceive	TX0006269552	11/18/2005
Tears of Pearl	TX0007040865	09/07/2009
A Terrible Beauty	TX0008405890	04/21/2017

B. Defendants

22. Defendant Apple Inc. ("Apple") is a California corporation headquartered at One Apple Park Way, Cupertino, CA 95014.

23. Defendant Craig Federighi is Apple's Senior Vice President of Software Engineering at Apple, oversees Apple's AI strategy including Apple Intelligence, and resides in Santa Clara County, California. Defendant Federighi was personally involved in the relevant decision making at Apple and directed, approved, or ratified the conduct that has led to widespread infringement.

24. Defendant John Giannandrea is Apple's Senior Vice President of Machine Learning and AI Strategy and resides in Santa Clara County, California. Defendant Federighi was personally involved in the relevant decision making at Apple and directed, approved, or ratified the conduct that has led to widespread infringement.

III. JURISDICTION AND VENUE

25. This Court has subject matter jurisdiction pursuant to 28 U.S.C. §§ 1331 and 1338(a), as this action arises under the Copyright Act of 1976, 17 U.S.C. § 101, *et seq.*

26. The Court has personal jurisdiction over the Defendants because the individual Defendants reside in this District and Apple is headquartered in Cupertino California, which is located in this District, where substantial copyright infringement occurred, including downloading and reproducing the copyrighted works. Defendants also marketed, sold, and distributed Apple Intelligence models from this District to California residents.

27. The Court also has personal jurisdiction over the Defendant because the Defendant has purposefully availed itself of the privilege of conducting business within this District.

28. Venue is proper under 28 U.S.C. §§ 1391(a) through (d) and 1400(a) because the individual Defendants reside in this District, Apple Inc. headquarters and principal place of business are located in Cupertino California, which is located in this District, Apple resides in this District, and substantial parts of the events or omissions giving rise to the claims occurred in or emanated from this District, including, without limitation, decisions made by Apple's governance and management personnel.

IV. DIVISIONAL ASSIGNMENT

29. Pursuant to Civil Local Rule 3-2(e), assignment to the San Jose Division is proper because Apple is headquartered in Cupertino California, the facts that give rise to this case occurred in Santa Clara County, and Senior Vice President Defendants both reside in Santa Clara County.

V. FACTUAL ALLEGATIONS

30. Founded by Steve Jobs and Steve Wozniak in 1976, Apple is one of the most recognizable brands in the world.[6] Apple is a multinational technology company recognized for revolutionizing the technology sector through its innovations of computer software, personal computers, mobile tablets, smartphones, and computer peripherals.[7] It is also recognized as the first to bring the graphical user interface (GUI) into mass adoption.[8]

[6] https://www.britannica.com/money/Apple-Inc/2011-present-Tim-Cook-as-CEO-and-the-first-trillion-dollar-company (last visited Oct. 22, 2025).
[7] *Id.*
[8] *Id.*

31. In 2024, Apple's revenue in 2024 was 390.8 billion.[9] Apple's domination is only growing, As of September 2025, Apple has a market cap of $3.473 Trillion USD, making Apple one of the world's most valuable companies.[10] In January 2025, Apple announced a quarter revenue of $124.3 billion, up 4 percent from a year ago.[11] In the announcement, Apple's CEO, Tim Cook, emphasized the importance of Apple's commercial AI platform, Apple Intelligence, stating "[w]e were thrilled to bring customers our best-ever lineup of products and services during the holiday season. Through the power of Apple silicon, we're unlocking new possibilities for our users with Apple Intelligence, which make apps and experiences even better and more personal."[12]

32. Apple trained its generative-AI models for Apple Intelligence using a vast data library that includes copyrighted works, such as books, texts, and images created by Plaintiff and the Class, copied without author consent, credit, or compensation.

33. A large language model ("LLM") is a software program trained by copying vast amounts of text known as the training dataset and feeding it into the model. During training, the model copies and analyzes expressive content from the dataset, adjusting its outputs to mirror the patterns in the text. After processing this material, the model can generate convincing simulations of natural language based on what it has learned. Apple Intelligence includes multi-modal models, meaning that it was trained on, can interpret, and may generate multiple modes of data, here, at least text and images.[13]

34. Apple is perceived to be behind its rivals in the development of its AI products, and like others in the tech industry, "was caught off guard" by the launch of ChaptGPT.[14] Its

---

[9] David Curry, *Apple Statistics (2025)*, BUSINESS OF APPS (Aug. 6, 2025) https://www.businessofapps.com/data/apple-statistics/.

[10] https://companiesmarketcap.com/apple/marketcap/ (last visited Oct. 22, 2025).

[11] Press Release, Apple Inc., Apple reports first quarter results (Jan. 30, 2025).

[12] *Id.*

[13] Harsh Shivam, *Inside Apple's new AI Models: How it works, where it gets its training data*, (July 22, 2025), https://www.business-standard.com/technology/tech-news/inside-apple-s-new-ai-models-how-it-works-where-it-gets-its-training-data-125072200594_1.html.

[14] Aaron Tilley and Wayne Ma, *Why Apple is Losing Ground in the AI Talent War (It's Not Just Money)*, The Information (July 22, 2025).

efforts to catch up has lead it to adopt a "measured approach to investing in AI," so far avoiding "the massive uptick in capital expenditures" incurred by Microsoft and others.[15] Notably, Apple "relies on other companies … to handle much of the spending."[16]

35. Apple avoids seeking licenses or paying authors for using their copyrighted works in training its AI models. Instead, it deliberately avoided compensation by relying on pirated book datasets already compiled by others and by scraping the web for content, including copyrighted works.

A. Defendants' AI Desperation Leads to Unscrupulous Behaviour.

36. AI is big business as the associated data center capital spend accounted for 92% of US growth in the first half of this year.[17] However, as the AI business opportunity became increasingly clear, it also became apparent that Apple was falling far behind.[18]

37. Responding to lackluster investor and customer sentiment, Apple launched a broad new marketing campaign as it introduced Apple Intelligence, with a dramatically enhanced AI Siri voice assistant as its centerpiece. At Apple's annual Worldwide Developers Conference, Apple senior management, including Defendant Federighi, made bold claims about Apple's AI capabilities.[19]

38. Apple's campaign worked, and when Apple Intelligence was unveiled in June 2024, featuring generative models and advanced tools, the company's value soared by over $200

[15] Anita Ramaswamy, *Apple's Low-Cost AI Strategy Wins Over Investors, Despite Risks*, The Information (October 28, 2024), https://www.theinformation.com/articles/apples-low-cost-ai-strategy-wins-over-investors-despite-risks?rc=gvxvc7.

[16] *Id*.

[17] Nick Lichtenberg, *Without data centers, GDP Growth was 0.1% in the first half of 2025, Harvard economist says*, FORTUNE (Oct. 7, 2025), https://fortune.com/2025/10/07/data-centers-gdp-growth-zero-first-half-2025-jason-furman-harvard-economist/.

[18] Samantha Murphy Kelly, *iPhone sales are plunging. Here's why*, CNN (Apr. 16, 2024), https://www.cnn.com/2024/04/15/tech/iphone-sales-plunging/index.html.

[19] Apple, WWDC 2024, YouTube (Jun. 10, 2024), https://www.youtube.com/watch?v=RXeOiIDNNek (quoted language at 1:10:45) ("Apple Intelligence is grounded in your personal information and context, with the ability to retrieve and analyze the most relevant data from across your apps, as well as to reference the content on your screen, like an email or calendar event you are looking at.").

billion, marking the most lucrative single day in the company's history and heralding a "new era" of Apple and the iPhone.[20]

39. But Apple's promises were entirely illusory. AI enhanced Siri never came. Apple's promised AI tools did not exist at the time Apple began promoting and selling them[21] and an AI enhanced Siri would not be available, if ever, until years later.[22] In a classic bait-and-switch form of false advertising, Apple lured consumers in by touting these core AI offerings and then sold consumers a product that failed to meet its grandiose promises.

40. Defendants' rapacious advertising behaviour directly coincides with their unscrupulous copying of Plaintiff's and the Class's copyrighted works. Defendants disregarded the rights of consumers in order to catch up with its competitors in AI, and they have similarly disregarded the rights copyright holders for the same reason.

B. Generative AI Tools and the Training Process.

41. AI is software designed to mimic human reasoning, creativity, or inference using algorithms based on statistical and mathematical models. Apple Intelligence includes multiple LLMs and text-to-image diffusion models developed and commercialized by Apple. LLMs are AI software designed to generate natural, human-like text responses to user prompts while text-to-image models take a natural language descriptions to create corresponding visual images.

42. Both LLMs and text-to-image diffusion models depend on large quantities of works, including those of Plaintiff and the Class. For example, even though Chat GPT 3 and 4 were built on the same general transformer framework, GPT-3 was trained on 175 million parameters and one-trillion words[23] while GPT-4 was trained on 1.76 trillion parameters and

---

[20] Ben Cohen, *Tim Cook on Why Apple's Huge Bets Will Pay Off*, WALL ST. J. (Oct. 20, 2024), https://www.wsj.com/style/tim-cook-interview-apple-intelligence-vision-pro-48c59018.

[21] Emily Forlini, *Apple Misled Consumers on the iPhone 16's AI Features, Report Finds*, PC MAG (Apr. 22, 2025), https://www.pcmag.com/news/apple-misled-consumers-on-the-iphone-16s-ai-features-report-finds.

[22] Lance Ulanoff, *This is what really happened with Siri and Apple Intelligence, according to Apple*, TECH RADAR (Jun. 10, 2025), https://www.techradar.com/computing/artificial-intelligence/this-is-what-really-happened-with-siri-and-apple-intelligence-according-to-apple.

[23] *GPT-3 vs. GPT-4*, GRAMMARLY, https://www.grammarly.com/blog/ai/gpt-3-vs-gpt-4/ (last visited Oct. 22, 2025).

approximately ten-trillion words to substantially increase performance.[24] The same is true for image generation as the influential LAION-5B open source database released in 2022, was fourteen times larger than the prior open-source data base of image-text pairings.[25] These large databases of images are what the image generators are trained on and act as "the fuel for AI Art."[26] This fuel for AI is the copyrighted works of Plaintiff and the Class.

43. The advancement of LLMs and image models has primarily been predicated on two major technical advancements, along with training on large quantities of high-quality data. LLMs were revolutionized by the transformer, the T in GPT ("Generative Pre-trained Transformer"), and text-to-image models were accelerated by diffusion.

44. In 2017, the transformer was introduced and provided the ability for LLMs to both process input sequences in parallel rather than sequentially and to weigh the importance of the different words or tokens within a section of text.[27] As transformer architecture progressed, developers began training their models on exponentially increasing volumes of text, which produced dramatically more effective responses. Roughly speaking, the progress of LLMs experienced over these past years has primarily been predicated on large quantities of valuable data and the transformer architecture.

45. Similarly, diffusion models are at the heart of image-generation[28] as they are trained by iteratively adding Gaussian noise, randomness, to images and then progressively denoising the data, here, pixels in an image.[29] Like LLMs, training diffusion models on

---

[24] Nisha Arya, *GPT-4 Details Have Been Leaked!*, KDNUGGETS (July 19, 2023), https://www.kdnuggets.com/2023/07/gpt4-details-leaked.html.

[25] Romain Beaumont, Laion-5B: *A New Era of Open Large-Scale Multi-Model Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/.

[26] *The Data Science Behind AI Art Generators*, THEDATASCIENTIST, https://thedatascientist.com/the-data-science-behind-ai-art-generators/ (last visited Oct. 22, 2025).

[27] Jacob Uszkoreit, *Transformer: A Novel Neural Network Architecture for Language Understanding*, GOOGLE RESEARCH (Aug. 31, 2017), https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/.

[28] Sascha Kirch, *The Rise of Diffusion Models – A new Era of Generative Deep Learning*, SCALEUP:AI (Mar. 27, 2024), https://towardsdatascience.com/the-rise-of-diffusion-models-a-new-era-of-generative-deep-learning-3ef4779f6e1b-2/.

[29] *Introduction to diffusion models for machine learning*, SUPERANNOTATE (Feb. 28, 2025), https://www.superannotate.com/blog/diffusion-models.

exponentially increasing quantities of valuable data dramatically increases the effectiveness of image generation. Generally, diffusion models trained on large volumes of quality data have been foundational for the recent advancement of image generation.

46. These large databases of text, images, and text-image pairs are called the "training dataset" or "training corpus," which is what the LLM or the image model learns from to generate the output. Apple Intelligence also consists of at least a multi-modal Foundation Model, which allows the model to better comprehend representations across modalities, such as the word "cat" and the pixels of a "cat" in an image. These multiple modes of training and generation provide the model with deeper context that increases the model's effectiveness.

47. A model is "pre-trained" when it copies and processes these texts or images to understand the statistical patterns, grammar, and relationships between words or relationships between pixels. This pre-training provides developers with "weights" or "parameters," which are a large set of numbers uniquely derived from the training data to better predict the next word that should occur in a sequence.

48. Once trained, the LLM uses its stored weights to generate natural language or image responses to user prompts. Each response is computed based on these weights, which imitates the protected expressions learned from the training data.

49. Much of Apple's high quality training data consists of copyrighted works, including books, text, and images created by Plaintiff and the Class, ingested and copied without their consent, credit, or compensation.

50. Stated otherwise, the training process consists of developers creating a "neural network"—a computational model that learns language patterns or the probabilistic assortment of pixel data from massive datasets.

51. These models are called neural networks because they consist of interconnected nodes ("neurons") arranged in layers, loosely resembling the structure of the human brain.

52. These "nodes" or "neurons" are the model's basic computational units, processing and transmitting fragments of information through the "brain" when the model performs a task.

53. The neural network learns by reading huge amounts of text and images and by paying attention to how words are used together or how pixels are arranged in an image. Over time, the model picks up patterns in data and language, like grammar, meaning, and how ideas are connected without anyone having to program those rules by hand.

54. Instead of being told what a sentence means, the model figures it out by studying lots of examples. As it learns, it builds a kind of mental map that shows how words and ideas relate to each other. This map is stored as numbers called "parameters." These numbers help the model understand, for example, that "dog" is more closely related to "puppy" than to "smartphone."

55. Once the model is trained, it uses this knowledge to respond to people by answering questions, writing text, completing tasks, or generating images.

56. Increasingly, artificial intelligence models are designed as multi-modal, here meaning that one single model is trained on both text and images. Apple Intelligence is multimodal as it can understand both text and image data.[30]

57. Apple Intelligence, like other models, requires an enormous amount of data to function. This scale is essential for identifying patterns, generalizing across topics, and producing high-quality, creative outputs. Training such a model uses trillions of words and millions or billions of images, some of which involve an unlicensed registered copyright.

58. Apple can measure its language training data in "tokens"—units of text that can represent a word, part of a word, or characters. On average, each word in the training data equals about 1.3 tokens.

59. To train an effective LLM, the data must be diverse in both subject matter and style. Without variety, including both long- and short-form content, the model performs poorly and has limited understanding. Apple's training data includes a broad range of sources such as scientific papers, code, and scraped web data. While books make up only a small portion

[30] *Apple Intelligence Foundation Language Models Tech Report 2024*, MACHINE LEARNING RESEARCH AT APPLE, https://machinelearning.apple.com/research/apple-foundation-models-tech-report-2025 (last visited Oct. 22, 2025).

compared to the much larger volumes of internet text and code, books, commonly recognized as high quality training data, are particularly important to effective LLM training.

60. After assembling the training corpus, Apple processes the data by removing duplicates, converting text into tokens, and randomizing sequences to prevent memorization. Tokens are then transformed into "vectors," which mathematically represent word relationships. The model's weights and biases, initially random, are adjusted through vast numbers of calculations as it learns language patterns.

61. This training creates a complex matrix capturing nuances like context (e.g., "bank" as a financial institution vs. "bank" as a river edge), grammar, and word structure. The process transforms random vectors into rich, contextual representations of language. The final model contains numerical values reflecting what it has learned but does not store or use the original training data once training is complete.

62. A pretrained model can function as an LLM, but its output quality and task relevance can be improved through "fine-tuning." This secondary training phase focuses the model on specific goals.

63. Inference is the process of using the LLM to generate outputs from user inputs such as spreadsheets, photos, or questions. Because the model is "probabilistic" rather than "deterministic," the same input may produce different outputs, and responses can't be precisely predicted.

64. For the reasons stated above, much of Apple's high quality training data consists of copyrighted works, including books by Plaintiff and the Class, copied without their consent, credit, or compensation.

65. For both the post- and pre-training processes in developing Apple Intelligence, Apple created multiple, unlicensed copies of the training data. As the U.S. Patent and Trademark

Office has observed, LLM "training" "almost by definition involve[s] the reproduction of entire works or substantial portions thereof." [31]

66. In sum, Apple trained its generative-AI models for Apple Intelligence using a vast data library that includes copyrighted works, such as books, text, and images created by Plaintiff and the Class, copied without author consent, credit, or compensation.

C. Apple Used Copyrighted Works to Train Its OpenELM Models.

67. In April 2024, Apple first announced the release of OpenELM on its website as a state-of-the-art open language model.[32]

68. The OpenELM language model family released in April 2024 includes OpenELM-270M, OpenELM-450M, OpenELM-1_1B, and OpenELM-3B, distinguished by parameter size. Open ELM-450M was trained with 450 million parameters while OpenELM-3B used 3 billion parameters, which enables more complex tasks but requires more computational power. By using the conventional nomenclature of identifying models based on their number of parameters, Apple implicitly highlights the importance of large volumes of quality data to the performance of their LLMs and other models.

69. Apple's OpenELM models are hosted on Hugging Face and include "model cards" that describe each model, its intended uses, limitations, training parameters, and dataset. The model cards state that the pre-training dataset includes a subset of the RedPajama dataset.

70. Apple further published a paper on OpenELM showing that much of its training data comes from the "Books" subset of the RedPajama dataset.[33] *See Figure 1* below.

[31] U.S. Patent & Trademark Office, Public Views on Artificial Intelligence and Intellectual Property Policy 29 (2020), available at https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf (last accessed Oct. 22, 2025).

[32] *OpenELM: An Efficient Language Model Family with Open Training and Inference Framework*, MACHINE LEARNING RESEARCH AT APPLE, https://machinelearning.apple.com/research/openelm (last visited Oct. 22, 2025).

[33] *See OpenELM Paper, supra* note 1.

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

Figure 1: Dataset used for pre-training OpenELM

71. Information about the RedPajama dataset is available on Hugging Face[34], where it is hosted, which until April 2024 stated that the dataset's "Books" component is a copy of the "Books3 dataset;" "Books" is automatically downloaded from Hugging Face when assembling RedPajama. Thus, anyone who used the "Books" subset to train an AI model used a copy of Books3.

72. Books3 is part of a separate AI training dataset called The Pile, curated by the research group EleutherAI. In December 2020, EleutherAI introduced The Pile in a paper titled "The Pile: An 800GB Dataset of Diverse Text for Language Modeling" ("The Pile Paper").[35] 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."

[34] *RedPajam-Data-1*, HUGGING FACE, https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T (last visited Oct. 22, 2025) ("The 'book' config is defunct and no longer accessible due to reported copyright infringement for the Book3 dataset contained in this config.").

[35] Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, ELEUTHER AI (Dec. 31, 2020), https://arxiv.org/pdf/2101.00027 at 1.

73. Bibliotik is a well-known "shadow library." Collections of pirated books, such as Library Genesis, Z-Library, and Bibliotik, are widely known to have circulated via the BitTorrent file-sharing network.[36] These sites have long attracted interest from the AI training community for their vast collections of highly valuable texts, which include extensive unlicensed copyrighted material.

74. Shawn Presser, who compiled the Books3 dataset, publicly confirmed that it is comprised of "all of Bibliotik" and includes approximately 196,640 books. These books in the Book3 dataset are stored as .txt files—a plain text format without formatting, fonts, or images—meaning the dataset contains the full text of all 196,640 books.[37]

75. Indeed, in an October 2020 Twitter thread, Presser explained how he created Books3.[38]

Figure 2: Explanation of the Contents of Books3

[36] Alex Reisner, *Revealed: The Authors Whose Pirated Books are Powering Generative AI*, THE ATLANTIC (Aug. 19, 2023) https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/ ("collections of pirated books, such as Library Genesis, Z-Library, and Bibliotik, that circulate via the BitTorrent file-sharing network.").

[37] Roger Montti, *Are ChatGPT, Bard and Dolly 2.0 Trained On Pirated Content?*, SEARCH ENGINE JOURNAL (Apr. 20, 2023), https://www.searchenginejournal.com/are-chatgpt-bard-and-dolly-2-0-trained-on-pirated-content/485089/ ("The Books3 dataset contains the text of books that were pirated and hosted at a pirate site called, bibliotic").

[38] *See* Shawn Pressor (@theshawwn), *Suppose you want to train a world-class GPT model, just like OpenAI*, TWITTER (Oct. 25, 2020, 1:32 AM), https://x.com/theshawwn/status/1320282149329784833?lang=en.

76. The Books3 dataset was available on Hugging Face until October 2023, when it was removed with a notice stating it was "defunct and no longer accessible due to reported copyright infringement."[39] Shawn Presser, creator of Books3, admitted they "almost didn't release the data sets at all because of copyright concerns."[40]

77. Prior to October 2023, using the "Books" subset of the RedPajama dataset required copying the Books3 dataset. By using the "Books" subset, and thereby the full text of each book in Books3, Apple copied entire works to train its OpenELM model.[41]

78. Apple's OpenELM models were thus trained on Books3, a known collection of pirated books, which included Plaintiff's and the Class's Infringed Works.

79. Even if Apple deletes Books3 from the dataset used to train OpenELM, its prior admitted reliance on the dataset makes certain that use of pirated books continues, as Books3 is already "deeply embedded in the very structure of the model's learned behaviors."[42]

80. Given that Plaintiff's and the Class's Infringed Works are included in Books3, Apple trained OpenELM on copies of these works, directly infringing the copyrights of Plaintiff and the Class.

---

[39] *RedPajam-Data-1*, Hugging Face, https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T (last visited Oct. 22, 2025).

[40] Kate Knibbs, The Battle Over Books3 Could Change AI Forever (Sept. 4, 2023) https://www.wired.com/story/battle-over-books3/.

[41] *See OpenELM Paper, supra* note 1, at 2.

[42] *See* John Funge, *Machine Unlearning: The Lobotomization of LLMs*, DARKREADING, (Feb. 26, 2025), https://www.darkreading.com/vulnerabilities-threats/machine-unlearning-lobotomization-llms ("[D]eleting data from a trained LLM is more like trying to retrieve a whole strawberry from a smoothie. Once the data has been integrated into the model, it's no longer a discrete chunk — it's deeply embedded in the very structure of the model's learned behaviors.") (last visited Oct. 22, 2025); *see also* A. Feder Cooper et al., *Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice*, THE GEN CENTER ET AL. (Dec. 9, 2024) at 2, https://arxiv.org/pdf/2412.06966 ("Even if one removed all in-copyright images of Spiderman from a model's training data, this does not mean it would be impossible for the model to generate outputs that resemble Spiderman when put to use. Generative-AI models are impressive in part because they are able to generate novel outputs that transcend the information that is exactly contained in their training data. It is therefore a mistake to think that making a limited set of targeted changes to a model's parameters is sufficient to make promises about what types of out-puts that model could or could not possibly produce [].").

D. Apple Intelligence Models Were Trained on Copyrighted Works.

81. On June 10, 2024, Apple issued a press release titled "Introducing Apple Intelligence, the personal intelligence system that puts powerful generative models at the core of iPhone, iPad, and Mac."[43] Apple Intelligence is the foundation model at the core of upgrades to Siri, system enhancements, and tools like Genmoji, Image Playground, and more.[44]

82. In July 2024, Apple released a report titled "Apple Intelligence Foundation Language Models" (the "FLM Report").[45]

83. The FLM Report details how two of its generative models, "AFM-on-device (AFM stands for Apple Foundation Model), a ~3 billion parameter language model, and AFM-server, a larger server-based language model—have been built and adapted to perform specialized tasks efficiently, accurately, and responsibly."[46]

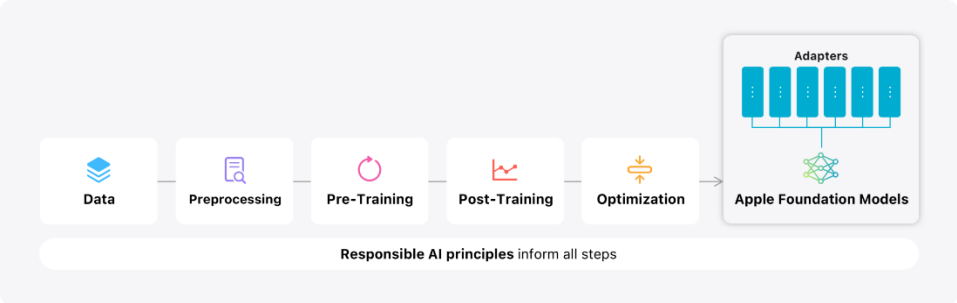

Figure 3: Modeling overview for the Apple Foundation models.[47]

84. A foundation model "is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks…"[48] Apple's FLM Report explains that "[w]riting is one of the most critical abilities for

[43] Press Release, Apple Inc., Introducing Apple Intelligence, the personal intelligence system that puts powerful generative models at the core of iPhone, iPad, and Mac (June 10, 2024).
[44] Press Release, Apple Inc., Apple Intelligence now features Image Playground, Genmoji, Writing Tools enhancements, seamless support for ChatGPT, and visual intelligence (Dec. 11, 2024).
[45] *Apple Intelligence Foundation Language Models*, APPLE INC. (Jul. 29, 2024), https://arxiv.org/html/2407.21075v1#S1.F1.
[46] *Id.*
[47] *Id.*
[48] Rishi Bommasani *et al.*, *On the Opportunities and Risks of Foundation Models*, CENTER FOR RESEARCH ON FOUNDATION MODELS (CRFM), STANDFORD INSTITUTE FOR HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (HAI) AND STANFORD UNIVERSITY (July 12, 2022) https://arxiv.org/pdf/2108.07258 at 3.

large language models to have, as it empowers various downstream use cases such as changing-of-tone, rewriting, and summarization."[49]

85. Of the two foundation language models – AFM-server and AFM-on-device – the AFM-server model is the larger model intended for use on the Apple cloud service, Private Cloud Compute while the AFM-on-device model is intended to be used directly on Apple devices such as iPhones, iPads, and laptops.[50]

86. The FLM Report reveals that both the AFM-server and AFM-on-device are trained on the same data by explaining, "before training AFM-on-device, we initialize it from a pruned 6.4B model (trained from scratch using the same recipe as AFM-server)."[51]

87. Apple describes three sources of training data in the FLM Report, which are: (1) data Apple "licensed from publishers"; (2) data "curated from publicly available or open-sourced datasets", and (3) "publicly available information crawled" by Apple's web-crawler, Applebot.[52] In a June 2025 article titled "Updates to Apple's On-Device and Server Foundation Language Models," Apple restates the same three categories of data for training.[53]

88. For the first source of its training data—data Apple "licensed from publishers"—Apple explains that it went "to lengths to identify and license a limited amount of high-quality data from publishers" and that "[t]hese datasets provide a natural resource of high quality long-context data, so we include them as part of the data mixture for the continued and context-lengthening pre-training."[54] Notably, Apple does not use this data set during "core pre-training" the main phase of training for the Foundation Language Models.[55]

89. For its second source of training data—data curated from publicly available or open-sourced datasets—Apple claims that it "evaluated and selected a number of high-quality

---

[49] *Apple Intelligence Foundation Language Models*, Apple (July 2024), https://arxiv.org/html/2407.21075v1.
[50] *Id.*
[51] *Id.*
[52] *Id.*
[53] *Updates to Apple's On-Device and Server Foundation Language Models*, Apple Inc. (June 9, 2024), https://machinelearning.apple.com/research/apple-foundation-models-2025-updates.
[54] *Apple Intelligence Foundation Language Models*, *supra* note 49.
[55] *Id.*

publicly-available datasets with licenses that permit use for training language models" and that it "filtered datasets to remove personally identifiable information before including them in the pre-training mixture."[56] Apple does not specify exactly what datasets it used in the FLM Report.

90. AI companies frequently claim that they get data to train their models from what is "publicly available" on the Internet.[57] However, "publicly available" simply means works that can be accessed and downloaded on the Internet. While "publicly available" sounds like "the company has permission to use the information … in many ways, it's more like the legal equivalent of 'finders, keepers.'"[58] It simply means that the AI company has not illegally hacked into a system.[59] The internet contains a massive amount copyrighted works by authors that have not granted a license or have otherwise given permission for reproduction of their works. The copying of such works is copyright infringement of pirated works.

91. Like Apple, Meta Platforms also trained its Llama language models on Books3. Meta described Books3 as a "publicly available dataset for large language models" even though none of the authors whose works appeared in Books3 ever consented to the inclusion of their works. The works in contained in Books3 were "publicly available" only in the sense that they could be accessed publicly, via the internet, without illegal hacking.

92. Additionally, Apple's claim that it uses data sets that are "open source" does not mean such works were made available with the permission of the author. In reality, copies of such works were made freely and publicly available by a third party, without the permission of the author.

93. In sum, Apple's use of publicly available or open-sourced datasets includes pirated works. Apple was in possession of a copy of Books3 for training of its OpenELM models. Upon information and belief, Apple's use of "publicly available or open-sourced datasets" includes Books3. As such, Apple used Books3 in the training dataset for its Foundation Language Models.

---

[56] *Id.*

[57] Ina Fried, *For AI firms, anything "public" is fair game*, AXIOS (Apr. 5, 2024), https://www.axios.com/2024/04/05/open-ai-training-data-public-available-meaning.

[58] *Id.*

[59] *Id.*

94. Apple has created an AI training data library containing "publicly-available or open-sourced datasets" – with pirated material – to train future models. Apple directly infringed on the copyrights of Plaintiff and Class Members because Plaintiff's Infringed Works were part of the Books3 dataset.

95. Finally, Apple uses training data obtained by Apple's web-crawler, Applebot. Apple states "we crawl publicly available information using our web crawler, Applebot."[60] Apple states that it "respect[s] the rights of web publishers to opt out of Applebot using standard robots.txt directives."[61] Further, it claims to "take steps to exclude pages containing profanity and apply filters to remove certain categories of personally identifiable information."[62] Notably, Apple does not state that it filters such information for copyrighted or pirated material.

96. Apple has been candid about its scraping of publicly available and often copyrighted content as the Senior Vice Presidents in charge of AI model training, Federighi and Giannandrea, have stated that Apple uses this publicly available internet data for training its models.[63]

97. For text and image data, Apple "continue[s] to source a significant portion of the pre-training data for our models from web content crawled by Applebot, spanning hundreds of billions of pages and covering an extensive range of language, locales and topics."[64] Apple focuses on "capturing high-fidelity HTML pages, which enrich the dataset with both text and structured metadata for aligning media with surrounding text content."[65]

98. Apple admits it "introduced image data into the pre-training pipeline leveraging high-quality licensed data along with publicly available image data. Using our web crawling

[60] *Apple Intelligence Foundation Language Models*, *supra* note 49.
[61] *Id.*
[62] *Id.*
[63] Andrew O'Hara, *Craig Federighi & John Giannandrea talk Apple Intelligence at WWDC*, APPLE INSIDER (Jun. 10, 2024), https://appleinsider.com/articles/24/06/10/craig-federighi-john-giannandrea-talk-apple-intelligence-at-wwdc.
[64] *Updates to Apple's On-Device and Server Foundation Language Models*, *supra* note 53.
[65] *Id.*

strategy, we sourced pairs of images with corresponding alt-texts."[66] Upon information and belief, such "publicly available" image data includes copyrighted content.

99. Many big web publishers have opted out of Apple's AI training, including Facebook, Instagram, Craigslist, Tumblr, The New York Times, The Financial Times, The Atlantic, Vox Media, the USA Today network, and Condé Nast, WIRED's parent company.[67] While Apple uses Applebot-Extended tag to allow sites to opt out of AI training, it still allows search indexing—"meaning that their pieces can still be included in Spotlight and Siri searches."[68]

100. Evidence of Apple's web crawler first appeared around November 2014.[69] In May 2015, Apple seemingly acknowledged their Applebot webcrawler with a webpage added to Apple's support site.[70] Apple's Foundation Language Models were very likely trained on such "publicly available" data before the publication of the FLM Report in July 2024, nearly ten years after Apple's web crawler first started scraping the internet. Thus, Apple's disclosure that it allows web publishers the option to opt-out came too late for such opt-outs to matter.

101. On information and belief, Apple has retained copies of scraped Applebot data in order to train future models. This includes data scraped before companies and web publishers were informed of the option or provided the opportunity to opt-out.

102. Further, Apple states that documents crawled by Applebot are "processed by a pipeline which performs quality filtering and plain text extraction."[71] More specifically, this includes "[s]afety and profanity filtering, using heuristics and model-based classifiers."[72]

---

[66] *Id.*

[67] Ben Lovejoy, *Many of the biggest websites have opted out of Apple Intelligence training*, 9TO5MAC (Aug. 29, 2024), https://9to5mac.com/2024/08/29/apple-intelligence-training-opt-outs/#:~:text=WIRED%20can%20confirm%20that%20Facebook,from%20Apple's%20AI%20training%20%5B%E2%80%A6%5D.

[68] *Id.*

[69] Neil Hughes, *Evidence shows Apple operating a mysterious Web crawling bot*, APPLEINSIDER (Nov. 6, 2024), https://appleinsider.com/articles/14/11/06/evidence-shows-apple-operating-a-mysterious-web-crawling-bot.

[70] Jennifer Slegg, *Apple Officially Acknowledges Their Applebot Web Crawler*, THE SEM POST (May 7, 2015), http://www.thesempost.com/apple-officially-acknowledges-their-applebot-web-crawler/.

[71] *Apple Intelligence Foundation Language Models*, *supra* note 49.

[72] *Id.*

103. "Classification models are a type of machine learning model that divides data points into predefined groups called classes. Classifiers are a type of predictive modelling that learns class characteristics from input data and learns to assign possible classes to new data according to those learned characteristics."[73] Thus, Apple's reference to "model-based classifiers" refers to a machine learning model trained to sort the data scraped by Applebot. On information and belief, these model-based classifiers are trained on datasets that include pirated, unlicensed, copyrighted works.

104. A 2024 report by George Wukoson, the lead attorney for Ziff Davis, and Joey Fortuna titled "The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLMs," revealed "that AI developers often favor high-quality content when selecting training data, especially content owned by premium media companies."[74] "Their findings showed that as datasets become more curated, the share of content from high-quality publishers rises significantly."[75] This material is often protected by registered copyrights, including works from leading publishers like The New York Times and News Corp.[76]

105. Further, the "publicly available information" Applebot has accessed and scraped includes so-called "shadow libraries," which operate outside traditional legal frameworks and distribute works without authorization from publishers or rights holders. Accordingly, Apple's training data set includes pirated material obtained from such "shadow libraries."

106. On information and belief, Apple intentionally obscures the training datasets it uses in developing its Apple Intelligence Foundational Language Models to hide the inclusion of copyrighted materials. It is commonly understood that "[m]any AI models are trained on

---

[73] Jacob Murel, *What are classification models?*, IBM (July 31, 2024), https://www.ibm.com/think/topics/classification-models.

[74] Rande Price, *AI developers favor premium media content for training*, DIGITAL CONTENT NEXT (Nov. 12, 2024), https://digitalcontentnext.org/blog/2024/11/12/ai-developers-favor-premium-media-content-for-training/; George Wukoson and Joey Fortuna, *The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLMs*, ZIFF DAVIS (Nov. 5, 2024), https://www.ziffdavis.com/wp-content/uploads/2024/11/The-Predominant-Use-of-High-Authority-Commercial-Web-Publisher-Content-to-Train-Leading-LLMs.pdf.

[75] *Id.*

[76] *Id.*

copyrighted material without permission."[77] Apple's decision to conceal the source of its training data is a result of prior public criticism, including reports that Apple utilized a subset of "The Pile" dataset containing video captions sourced from thousands of YouTube videos without authorization from the creators.[78]

107. Apple trained its Apple Intelligence models on "curated publicly available or open-sourced datasets," which contained copyright materials—including works owned by Plaintiff. This practice is consistent with Apple's approach to training its OpenELM model, for which Apple publicly acknowledged the use of datasets such as The Pile, referring to them as "public datasets." Describing these datasets as "public" does not mean the contents were lawfully acquired or that the dataset provider had the authority to grant a valid license for the use of the underlying copyrighted works.

108. A growing body of publicly disclosed licensing arrangements reflects the emergence of structured markets for AI training data. Further, organizations have developed licensing systems, such as the Copyright Clearance Center's collective licensing framework and the Created by Humans platform, which aim to facilitate ethical and lawful access to copyright content for AI Development. Additionally, several dataset licensing entities have formed an industry consortium known as the Dataset Providers Alliance. The value of the AI training dataset market was estimated to be approximately $2.62 billion in 2024 and is projected to reach $18.47 billion in 2034.[79] These developments underscore the commercial significance of licensing copyrighted materials for AI training and the increasing recognition of the need for enforceable rights and transparent licensing practices.

[77] *Should AI Models Disclose Their Training Data?*, EDUWIK (Aug. 7, 2025), https://eduwik.com/should-ai-models-disclose-their-training-data/.

[78] *See* Omar Gallaga, *Apple Trained Its AI on YouTube Transcripts Without Permission, Report Says*, CNET (July 17, 2024), https://www.cnet.com/tech/services-and-software/apple-trained-its-ai-on-youtube-transcripts-without-permission-report-says/.

[79] Press Release, Research and Markets, AI Training Dataset Market Report 2025: Market to Reach $18.47 Billion by 2034 from $2.6 Billion in 2024, E-Commerce Expansion and LLM Adoption, Despite Talent Shortage Risks (May 12, 2025).

109. Apple and its industry competitors recognize that there is a market for paying creators to use their works for AI training and understand the value of such copyrighted works. For example, Apple, Amazon, Meta, and Google have individually struck deals with Shutterstock to allow them to train their AI on millions of images in the Shutterstock library.[80] These deals reportedly ranged from $25 million to $50 million dollars each.[81]

110. In 2023, it was reported that Apple opened negotiations with major news and publishing organizations to use their material in their AI development.[82] According to the New York Times, "[t]he technology giant has floated multiyear deals worth at least $50 million to license the archives of news articles."[83] "The news organizations contacted by Apple include Condé Nast, publisher of Vogue and The New Yorker; NBC News; and IAC, which owns People, The Daily Beast and Better Homes and Gardens."[84]

111. Furthermore, Apple is considering introducing a paid subscription tier for its Apple Intelligence products.[85] According to reports, Apple could charge its users up to $20 for its advanced artificial intelligence features to boost growth of its lucrative services business.[86] Apple is already profiting by licensing its foundational AI models to developers, who were granted access to a beta model via the Apple Developer Program, which requires a $99 annual subscription.[87] It is notable that the cost of this subscription "might be cheaper than paying to

---

[80] Tom Pritchart, *Apple has licensed millions of Shutterstock images to train its AI – what this means for the iPhone*, TOM'S GUIDE (April 8, 2024), https://www.tomsguide.com/ai/ai-image-video/apple-has-licensed-millions-of-shutterstock-images-to-train-its-ai-what-this-means-for-your-iphone.

[81] *Id.*

[82] Benjamin Mullin and Tripp Mickle, *Apple Explores A.I. Deals With News Publishers*, N.Y. TIMES (Dec. 22, 2023), https://www.nytimes.com/2023/12/22/technology/apple-ai-news-publishers.html.

[83] *Id.*

[84] *Id.*

[85] Arjun Kharpal, *Apple could charge up to $20 for some Apple Intelligence AI features, analysts say*, CNBC (Aug. 8, 2024), https://www.cnbc.com/2024/08/08/apple-could-charge-20-for-some-apple-intelligence-features-analysts.html.

[86] *Id.*

[87] Emily Forlini, *Apple Intelligence Is About to Face Its Toughest Test: Software Developers*, PC Mag (June 11, 2025), https://www.pcmag.com/opinions/apple-intelligence-faces-toughest-test-software-developers-wwdc-2025.

access models from other brands."[88] Apple Intelligence and AI generally present Apple with massive business opportunities built upon the Infringed Works of Plaintiff and the Class.

E. Apple's Actions Impaired the Market for Plaintiff's and Class Members' Works.

112. In order to train its AI models, Apple downloaded, scraped, and copied large quantities of copyrighted material—including unlawfully complied datasets, including Books3—that encompass Plaintiff's works.

113. Apple did not have permission from Plaintiff to use her Infringed Works nor did Apple pay Plaintiff for use of her works. In bypassing the proper licensing or purchase process, Apple has deprived Plaintiff of the revenues she would have realized through direct negotiations with Plaintiff or their authorized agents. Moreover, the creation of private libraries drawn from these illicitly assembled datasets for AI training purposes threatens to depress sales and harm the established market for access to Plaintiff's works.

114. Apple's unauthorized use of Plaintiff's Infringed Works, has caused and continues to cause grievous harm to the market for Plaintiff's works. Plaintiff and Class Members license their work for commercial use. Apple's conduct undermines the emergence of lawful licensing regimes and has disrupted the market for Plaintiff's works by accessing and taking advantage of the unlawful pirating of such works. As a result, Plaintiff has faced and continues to face harms including, but not limited to, lost sales and recognition for their works.

115. Apple trained its AI models on verbatim copies of Plaintiff's Infringed Works so that the models can generate content that displaces the work that Plaintiff is paid to produce. This further diminishes the demand for books, text, visual art, and original, human-produced content.

F. Apple's Piracy Is Not Fair Use.

116. It is near impossible that "any accused infringer could ever meet its burden of explaining why downloading source copies from pirate sites *that it could have purchased or otherwise accessed lawfully* was itself reasonably necessary to any subsequent fair use." *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1025 (N.D. Cal. 2025).

[88] *Id.*

117. Defendant's verbatim copying of troves of copyrighted work to avoid compensating rights holders is not and cannot be fair use. The piracy that defendant engaged in is infringement without a defense. It strains reason to believe that a defendant could satisfy the burden of fair use when stealing and copying at least thousands of copyrighted works, regardless of any subsequent use. Pirating otherwise purchasable works is copyright infringement.

118. Apple's piracy was intended to build a central library that Defendant could have paid for, but chose not to. Apple did not copy from authorized purchased versions of copyrighted work, but instead downloaded thousands of unauthorized copies from a library known to be from an illicit torrent and used Applebot to indiscriminately hoover up countless copyrighted works for its centralized commercial database, all with complete disregard to the rights of copyright holders.

119. Apple, which touts itself as the technology company that values privacy and security, used these copyrighted works not only to train OpenELM and Apple Intelligence, but also to create a centralized library for any number of future purposes. Apple built this unauthorized database as a substitute for paid copies of copyrighted works. Any subsequent use of this library, such as to train an LLM, an image model, or a multi-modal model, is contingent and necessarily predicated on the creation of this highly valuable centralized database. Creating this library of copyrighted materials is its own use, is not transformative, and is indefensible infringement.

120. Moreover, any consideration for intermediate copying to later build an LLM is undermined by Defendant's acts of piracy and by the fact that Defendant maintained its libraries through multiple iterations of Apple's machine learning models. Defendant's use of Plaintiff's works was not simply an intermediate step, but was a commercially exploitative end unto itself. Arguments of intermediate copying cannot save Defendants because Plaintiff's works "were acquired and retained, as a central library." *Id.* at 1026.

121. The exploitation of Plaintiff's works was not indirect, it was and is a direct download and scraping of the entirety of Plaintiff's works into a centralized database with no transformation of form. Plaintiff's works were copied and maintained for future commercial

purposes. Defendants' copying was not transitory as Plaintiff's works were not immediately destroyed, but were retained to train multiple of Apple's AI models, generating billions of dollars for Apple. Apple's "piracy of otherwise available copies is inherently, irredeemably infringing even if the pirated copies are immediately used for [a] transformative use and immediately discarded." *Id.* at 1025.

122. Piracy is piracy, but even if the court applies the statutory fair use factors to Apple's unmitigated copying through Books3 and its extensive web scraping, these factors weigh against fair use. The four factors of fair use are: (1) "the purpose and character of the use," (2) "the nature of the copyrighted work," (3) "the amount and substantiality of the portion used," and (4) "the effect of the use upon [Plaintiff's] potential market." 17 U.S.C. § 107.

123. **Factor one.** The purpose and character of Apple's piracy is not a close call, as its use is strictly commercial and it is not transformative. The factor one analysis includes (1) whether the purpose of copying is commercial or non-commercial and (2) whether the purpose is transformative. Apple's piracy was and is not for educational or commentary purposes, but furthered a commercial endeavor to aggregate a massive library of valuable copyrighted works that helped add billions to Apple's market cap. Further, there is nothing transformative about pirating entire verbatim copies of Plaintiff's works to source free market substitutes for an unauthorized commercial database. Piracy is not transformative and is not fair use.

124. **Factor two.** The factor two analysis primarily questions whether the nature of the copied work is more creative or factual, and secondarily questions whether the copied works are accessible or out of print. Plaintiff's works are highly creative and are therefore entitled to broader protections under copyright. Further, inaccessible works are provided less protection because there is no market for inaccessible works that a defendant's copying could displace. However, "this is not a case where source copies were unavailable," *Anthropic PBC*, 787 F. Supp. 3d at 1027, and Apple chose not to purchase or license this copyrighted material. The nature of Plaintiff's Infringed Works ensures Plaintiff the greatest copyright protection and narrows Defendant's argument for fair use.

125. **Factor three.** The third factor asks whether a qualitatively and/or quantitatively substantial portion of Plaintiff's Infringed Works were stolen. The entirety of Plaintiff's Infringed Works was taken, verbatim reproduced, and stored for Apple's use on an ongoing basis. These works were not copied in a modified form, such as an image thumbnail, such that the substance was not taken; far more than small snippets of these books were copied. Plaintiff's Infringed Works were copied in their entirety to take both the whole and the heart of the work. The third factor weighs strongly in favor of infringement.

126. **Factor four.** The harm to Plaintiff's market is clear as "[t]he copies used to build a central [training] library *and* that were obtained from pirated sources plainly displaced demand for Authors' books — copy for copy." *Id.* at 1033. Factor four considers whether Apple's actions displaced or diluted the market for Plaintiff's books. More accurately, market displacement considers both whether the infringer's conduct itself causes harm or whether widespread replication of the infringer's conduct would harm Plaintiff's market. At every turn, Apple's conduct injures Plaintiff. Apple injures Plaintiff's market when downloading a pirated library or scraping the web for text-image pairs, when assembling this centralized database for ongoing commercial use, and even when training its various LLMs.

127. Apple's unauthorized Books3 download and web scraping displaced Plaintiff's market through Defendant's own piracy. Widespread piracy inevitably leads to market harm as it is difficult for creators that charge for their work, *i.e.*, through book sales and licensing, to compete with free. Moreover, while merely pointing towards a theoretical licensing market is insufficient, Apple has demonstrated that there is a thriving licensing market for Plaintiff's Infringed Works as Apple has paid millions to rights holders it chose not to infringe upon. Apple's individual actions have injured Plaintiff's market and widespread adoption of Apple's tactics will only exacerbate this harm.

128. Even when considering a subsequent purpose for Defendant's activity, training its AI models, Plaintiff and the Class are severely injured. After Apple copies Plaintiff's Infringed Works to train its models, these tools are then capable of generating functionally limitless books, text, or images at an infinitesimal fraction of the time and cost. This makes AI training unlike any

preceding case of infringement as Apple then generates literally millions of subsequent works that supplant Plaintiff's markets. Even if Defendant's outputs may not have a statistically high likelihood of generating substantially similar copies of Plaintiff's Infringed Works, Apple's AI tools flood the market for sufficiently similar works that have a pronounced effect on Plaintiff and the Class. This is not hypothetical or speculative, this substitution is presently occurring as Class Members are already facing displacement due to Apple's infringement.[89] As the Supreme Court recently indicated, even if two works are not perfect substitutes, so long as market displacement would occur through widespread adoption of Defendants' copying, the infringement is unfair. *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508 (2023).

129. The purpose of copyright, since the beginning of this country, is to incentivize original creativity. Apple's piracy and resultant AI models ensure that consumers can generate sufficiently superseding work in mere seconds. This displacement and dilution of the market for the works of Plaintiff and the Class severely disincentivizes their future creativity. Within the context of copyright's historic purpose, Apple's market displacement requires an adjudication that Apple's conduct is not fair use.

VI. CLASS ACTION ALLEGATIONS

130. Plaintiff brings this action individually and on behalf of the Class, as defined below, pursuant to Rules 23(a), 23(b)(2) and 23(b)(3), of the Federal Rules of Civil Procedure:

> All legal or beneficial owners of a registered U.S. copyright in any work that has been or is being used by Apple in the compilation or creation of training datasets or in the training, research, or development of any of Apple's AI models during the Class Period. For purposes of this definition, copyrighted works are limited to those registered with the United States Copyright Office within five years of the work's publication before being trained on by Apple, or within three months of publication.

[89] Xiang Hui and Oren Reshef, *Is Generative AI a job killer? Evidence from the freelance market* (July 8, 2025), https://www.brookings.edu/articles/is-generative-ai-a-job-killer-evidence-from-the-freelance-market/ ("[F]reelancers in occupations more exposed to generative AI have experienced a 2% decline in the number of contracts and a 5% drop in earnings since the release of new AI software in 2022.").

131. Plaintiff reserves the right to amend the class definition and/or subclass definition. Certification of Plaintiff's claims for class wide treatment is appropriate because Plaintiff can prove the elements of her claims on a class wide basis using the same evidence as would be used to prove those elements in individual actions alleging the same claims.

132. The "Class Period," as defined herein, extends from at least three years prior to the filing of this Complaint through the present.

133. The Class excludes: (a) Defendant; (b) Defendant's parent companies, subsidiaries, and affiliates; (c) Defendant's officers, directors, management, employees, and agents; (d) all governmental entities; and (e) the judges assigned to this case, their chambers staff, and members of their immediate families.

134. **Numerosity and Ascertainability:** The Class is comprised of thousands of authors and copyright holders, making joinder of all members impracticable. The identities of Class members can be readily determined from Defendant's business records and, at a minimum, from the contents of the Books3 database that Apple unlawfully downloaded.

135. The Class is comprised of thousands of authors and copyright holders, making joinder of all members impracticable. The identities of Class members can be readily determined from Defendant's business records and from the contents of the Books3 database that Apple unlawfully downloaded.

136. The Class is comprised of at least thousands of authors and copyright holders, making joinder of all members impracticable. The identities of Class members can be readily determined from Defendant's business records and, at a minimum, from the contents of the Books3 database that Apple unlawfully downloaded.

137. **Commonality and Predominance:** This action involves common questions of law and fact which predominate over any question solely affecting individual Class members. These common questions include:

a. Whether Plaintiff's and Class Members' works were included in the training datasets Apple used to develop its Apple Intelligence product, including the RedPajama datasets;

b. Whether Apple's inclusion of Plaintiff's and Class Members' works in its training datasets constituted, or required, reproduction of those works by Apple;

c. Whether Apple reproduced Plaintiff's and Class Members' works without authorization;

d. Whether Apple infringed the copyrights of Plaintiff and Class Members by scraping or downloading their Infringed Works and other copyrighted works and using them in the compilation of their training datasets;

e. Whether Apple infringed the copyrights of Plaintiff and Class Members by scraping or downloading their Infringed Works and other copyrighted works and using them in its OpenELM or Apple Intelligence models or products;

f. Whether the Court should enjoin Defendant from continuing the unlawful conduct alleged herein;

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

h. Whether Apple's infringement was willful.

138. These common questions of law and fact predominate over questions that affect only individual Class members.

139. **Typicality:** Plaintiff's claims are typical of the other Class members' claims because all Class members were comparably injured through Defendant's substantially uniform misconduct, as described above. Plaintiff is advancing the same claims and legal theories on behalf of themselves and all other members of the Class that they represent, and there are no defenses that are unique to the Plaintiff. The claims of Plaintiff and Class members arise from the same operative facts and are based on the same legal theories.

140. **Adequacy:** Plaintiff will fairly and adequately represent and protect the interests of the Class, have no interests incompatible with the interests of the Class, and have retained counsel competent and experienced in class action litigation.

141. **Superiority:** Class treatment is superior to other options for resolution of the controversy because the relief sought for each Class Member is small, such that, absent

representative litigation, it would be infeasible for Class Members to redress the wrongs done to them. Allowing the claims to proceed on a class basis will eliminate the possibility of repetitive litigation. Defendant has acted on grounds applicable to the Class, thereby making appropriate final injunctive and declaratory relief concerning the Class as a whole.

142. Risk of Prosecuting Separate Actions: Separate actions by individual Class Members would risk inconsistent adjudications and inefficiently consume limited judicial resources.

143. Finally, at the very minimum, there are multiple common issues relating to Apple's uniform context, such as (but not limited to) their ingestion, reproduction, and willfulness.

VII. CAUSES OF ACTION

COUNT ONE
DIRECT INFRINGEMENT
17 U.S.C. § 501
(On Behalf of the Plaintiff and the Nationwide Class)

144. Plaintiff reallege and incorporate by reference paragraphs 1-143 as if fully set forth herein.

145. Plaintiff and members of the proposed Class created the original works and own the registered copyrights of those works that Apple copied and appropriated to train its artificial intelligence models.

146. Plaintiff's and members of the proposed Class thus hold the exclusive rights, including the rights of reproduction, distribution, and the creation of derivative works under 17 U.S.C. § 106.

147. Apple infringed and continues to infringe on Plaintiff's and Class Members' exclusive rights under 17 U.S.C. § 106 by reproducing the works owned by Plaintiff and Class Members' to train its artificial intelligence models without permission, consent, or license.

148. Apple copied works sourced from shadow libraries containing pirated or unlawfully acquired works, which infringed on Plaintiff's and Class Members' exclusive rights.

149. Using Applebot, Apple scraped the internet by systematically copying large volumes of web pages to train its AI systems.

150. The web pages scraped by Apple contained pirated or unlawfully acquired works, which infringed on Plaintiff's and Class Members' exclusive rights.

151. On information and belief, Apple's conduct was and is willful because Apple is infringing on the exclusive rights of Plaintiff and Class Members knowing that the works are copyrighted and is profiting from mass copyright infringement.

152. Apple's infringement has directly resulted and is directly causing ongoing monetary harm, including, but not limited to (1) lost licensing revenue that Plaintiff and Class members would have received had Apple properly obtained authorization and (2) loss of value to Plaintiff and Class Members works and portfolios due to Apple's mass infringement that has replaced and depressed the overall market for the Plaintiff's and Class Members' respective commercial markets.

153. Plaintiff and Class Members are entitled to all remedies available under the Copyright Act, including, but not limited to, statutory damages, actual damages, injunctive relief, restitution of profits, disgorgement, and attorneys' fees and costs.

154. Plaintiff and Class members have been and continue to be irreparably injured due to Apple's conduct. Apple will continue to infringe on Plaintiff and Class Members' exclusive rights unless its infringing activity is enjoined by this Court. Plaintiff and Class Members are thus entitled to permanent injunctive relief barring Apple's ongoing infringement.

VIII. PRAYER FOR RELIEF

155. WHEREFORE, Plaintiff, individually and on behalf of all others similarly situated, seeks judgment against Defendant, as follows:

a) For an Order certifying this action as a Class Action and appointing Plaintiff as Class Representative and her counsel as Class Counsel;

b) For equitable relief enjoining Defendant from engaging in the wrongful conduct complained of herein;

c) For an award of actual damages, treble damages, compensatory damages, statutory damages, nominal damages, and/or statutory penalties in an amount to be determined, as allowable by law;

d) For disgorgement and restitution of all earnings, profits, compensation, and benefits received by Defendants as a result of their unlawful acts, omissions, and practices;

e) For an award of punitive damages, as allowable by law;

f) For an award of attorneys' fees, as allowable by law;

g) Pre- and post-judgment interest on any amounts awarded; and

h) Such other and further relief as this Court may deem just and proper.

IX. DEMAND FOR TRIAL BY JURY

Plaintiff demands a trial by jury of all issues so triable.

Dated: October 22, 2025

COTCHETT, PITRE & MCCARTHY, LLP

By: */s/ Joseph W. Cotchett*

Joseph W. Cotchett, Cal. Bar No. 36324
Brian Danitz, Cal. Bar No. 247403
Gia Jung, Cal. Bar No. 340160
Caroline Yuen, Cal. Bar No. 354388
840 Malcom Road
Burlingame, CA 94010
Tel: 650-697-6000
Fax: 650-697-0577
jcotchett@cpmlegal.com
bdanitz@cpmlegal.com
gjung@cpmlegal.com
cyuen@cpmlegal.com

Karin B. Swope (*pro hac vice* to be filed)
Thomas E. Loeser, Cal. Bar No. 202724
Andrew Fuller (*pro hac vice* to be filed)
Jacob M. Alhadeff (*pro hac vice* to be filed)
1809 7th Ave., Ste. 1610
Seattle, WA 98101
Tel: (206) 802-1272
Fax: (206) 299-4184
kswope@cpmlegal.com
tloeser@cpmlegal.com
afuller@cpmlegal.com

jalhadeff@cpmlegal.com

BLEICHMAR FONTI & AULD LLP

By: */s/ Lesley E. Weaver*
Lesley E. Weaver (SBN 191305)
Anne K. Davis (SBN 267090)
Joshua D. Samra (SBN 313050)
1330 Broadway, Suite 630
Oakland, CA 94612
Tel. (415) 445-4003
lweaver@bfalaw.com
adavis@bfalaw.com
jsamra@bfalaw.com

Counsel for Plaintiff