Joseph R. Saveri (SBN 130064)
Christopher K.L. Young (SBN 318371)
William W. Castillo Guardado (SBN 294159)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
cyoung@saverilawfirm.com
wcastillo@saverilawfirm.com

Lesley E. Weaver (SBN 191305)
Anne K. Davis (SBN 267090)
Joshua D. Samra (SBN 313050)
STRANCH, JENNINGS & GARVEY, PLLC
1111 Broadway, Suite 03-130
Oakland, CA 94607
Tel. (341) 217-0550
Email: lweaver@stranchlaw.com
adavis@stranchlaw.com
jsamra@stranchlaw.com

Karin B. Swope (*pro hac vice*)
Thomas E. Loeser, (SBN 202724)
Andrew Fuller (*pro hac vice* to be filed)
Jacob M. Alhadeff (*pro hac vice* to be filed)
COTCHETT, PITRE & MCCARTHY, LLP
1809 7th Ave., Ste. 1610
Seattle, WA 98101
Tel: (206) 802-1272
Fax: (206) 299-4184
Email: kswope@cpmlegal.com
tloeser@cpmlegal.com
afuller@cpmlegal.com
jalhadeff@cpmlegal.com

Attorneys for Individual and Representative Plaintiffs E. Molly Tanzer, Jennifer Gilmore, Tasha Alexander, Jon McGoran, and Art Kleiner

[Additional counsel on signature block below]

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

E. MOLLY TANZER, an individual; JENNIFER GILMORE, an individual; TASHA ALEXANDER, an individual; JON MCGORAN, an individual; and ART KLEINER, an individual.

Individual and Representative Plaintiffs,

v.

SALESFORCE, INC.

Defendant.

Master File Case No. 25-cv-08862-CRB
Consolidated with Case No. 3:25-cv-09560

FIRST AMENDED CONSOLIDATED COMPLAINT

JURY TRIAL DEMANDED

CLASS ACTION

TABLE OF CONTENTS

Plaintiffs E. Molly Tanzer, Jennifer Gilmore, Tasha Alexander, Jon McGoran, and Art Kleiner (together "Plaintiffs"), on behalf of themselves and all others similarly situated, bring this first amended consolidated complaint ("FACC") against Defendant Salesforce, Inc. ("Salesforce").

OVERVIEW

1. Defendant Salesforce, Inc. copied and pirated hundreds of thousands of copyrighted books to develop its XGen series of large language models ("LLMs"). The training dataset for these models—described as "legally compliant" by Salesforce—consists of the notorious RedPajama and The Pile datasets that contain copies of these unlawfully-obtained copyrighted books. Salesforce unlawfully downloaded, stored, copied, and used these datasets to develop its XGen series of LLMs.

2. Both RedPajama and The Pile contain the Books3 corpus, which contains hundreds of thousands of copyrighted books that were acquired without the authorization or consent of the authors.

3. Plaintiffs and Class members are copyrighted authors. They own registered copyrights in books that were included in the RedPajama and The Pile datasets that Defendant downloaded, copied, stored, and used without their permission or compensation.

4. Plaintiffs and Class members never authorized Defendant to download, copy, store, or use their copyrighted works. Defendant has never compensated Plaintiffs and Class members for downloading, copying, storing, or using their copyrighted works.

5. Salesforce has and continues to benefit commercially from its massive acts of copyright infringement. It does so by securing lucrative contracts with enterprise customers for the use of its LLMs, including through the Agentforce AI platform.

6. Through the above acts, Defendant has infringed Plaintiffs' copyrighted works and continues to do so by continuing to store, copy, use, and process the datasets containing copies of Plaintiffs' and the putative Class's copyrighted books.

JURISDICTION AND VENUE

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

8. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Defendant is headquartered in this district. Salesforce created the XGen series of large language models. Defendant distributes these models commercially. Therefore, a substantial part of the events giving rise to the claim occurred in this District. A substantial portion of the affected interstate trade and commerce was carried out in this District. Defendant has transacted business, maintained substantial contacts, and/or committed overt acts in furtherance of its unlawful acts throughout the United States, including in this District. Defendant's conduct has had the intended and foreseeable effect of causing injury to persons residing in, located in, or doing business throughout the United States, including in this District.

9. 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.

PARTIES

A. PLAINTIFFS

10. Plaintiff E. Molly Tanzer is an author who lives in Colorado. Ms. Tanzer owns registered copyrights in multiple books, including *Creatures of Will and Temper*.

11. Plaintiff Jennifer Gilmore is an author who lives in Pennsylvania. Ms. Gilmore owns registered copyrights in multiple books, including *If Only*.

12. Plaintiff Tasha Alexander is an author and the pseudonym for Anastasia Grant, who lives in Wyoming. Ms. Alexander owns registered copyrights in multiple books, including *And Only to Deceive*, *Tears of Pearl*, and *A Terrible Beauty.*

13. Plaintiff Jon McGoran is an author who lives in Elkins Park, Pennsylvania. Mr. McGoran owns registered copyrights in multiple books, including *Spliced*.

14. Plaintiff Art Kleiner is an author who lives in New York, NY. Mr. Kleiner owns registered copyrights in multiple books, including *The Age of Heretics: Heroes, Outlaws, and the Forerunners of Corporate Change.*

15. A non-exhaustive list of registered copyrights owned by Plaintiffs that were copied and pirated by Defendant is included as **Exhibit A**.

B. DEFENDANT

16. Defendant Salesforce, Inc. is a Delaware corporation with its principal place of business at 415 Mission St, 3rd Fl, San Francisco, CA 94105.

C. AGENTS AND CO-CONSPIRATORS

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

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

FACTUAL ALLEGATIONS

A. "The Pile" and "RedPajama" Contain Plaintiffs' and Class Members' Works.

19. In October 2020, the Books3 dataset, consisting of approximately 196,640 books sourced from the online pirate "shadow library" Bibliotik, was released for free online to provide AI developers with access to high-quality training data. The Books3 dataset included Plaintiffs' copyrighted books without their permission in violation of the U.S. Copyright Act.

20. The Pile is a dataset curated by a research organization called EleutherAI for use in training AI models. In December 2020, EleutherAI introduced this dataset in a paper called "The Pile:

An 800GB Dataset of Diverse Text for Language Modeling."[1] The paper provides a description of the Books3 dataset contained within The Pile:

> 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 (BookCorpus2). We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.[2]

21. Bibliotik is one of a number of notorious pirate websites that also includes Library Genesis (aka LibGen), Z-Library (aka B-ok), Sci-Hub, and Anna's Archive. These pirate libraries have long been of interest to the AI-training community because they host and distribute vast quantities of unauthorized copyrighted material, including books. For that reason, these pirate libraries also violate the U.S. Copyright Act.

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

23. EleutherAI's website encourages the public to download The Pile dataset from a website known as "The-Eye." Anyone who downloads The Pile from The-Eye is therefore also downloading a copy of Books3.

24. The Pile is also available for download from the "Hugging Face" website. Before October 2023, the Books3 subset of The Pile was available for download from Hugging Face as a standalone dataset. But in October 2023, the Books3 dataset was removed with a message that it "is defunct and no longer accessible due to reported copyright infringement."[3]

25. Books3 has also been included within another dataset known as RedPajama created by the company Together AI. Released in or around April 2023, the RedPajama dataset contained a subset

[1] Leo Gao et al., *The Pile: An 800 GB Dataset of Diverse Text for Language Modeling*, (Dec. 31, 2020), https://arxiv.org/pdf/2101.00027.pdf (on file with Joseph Saveri Law Firm, LLP)

[2] *Id.* at 3–4.

[3] *Data Sets: The Pile: Books3*, Hugging Face, https://web.archive.org/web/20231127101818/https://huggingface.co/datasets/the_pile_books3 (archived Nov. 27, 2023).

called "Books" or "RedPajama-Books" that was a direct copy of the Books3 dataset. The RedPajama dataset is available for download from Hugging Face. Before October 2023, the Books3 subset was "downloaded from Huggingface [sic]" when a user ran the script that automatically assembled the RedPajama dataset.[4] After the Books3 dataset was removed from Hugging Face in October 2023, the RedPajama dataset documentation similarly added a message that Books3 is defunct "due to reported copyright infringement."[5]

26. But before October 2023, anyone who downloaded the RedPajama or The Pile datasets from Hugging Face was necessarily downloading a copy of the Books3 dataset.

27. Plaintiffs' copyrighted books listed in Exhibit A are among the works in the Books3 dataset. Below, these books are referred to as the **Infringed Works**.

B. AI Models Require Datasets.

28. A large language model ("LLM") is AI software designed to emit convincingly naturalistic text outputs in response to user prompts.

29. Though an LLM is a software program, it is not created the way most software programs are—that is, 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 input material is called the *training dataset*.

30. Training consists of a multi-stage process (known as the training pipeline) that includes the acquisition and curation of the dataset, processing of the dataset, feeding the dataset into the model so that the model can extract the patterns and relationships from the protected expression contained therein; and further fine-tuning the model for more specialized uses with even more data. This process also involves experimentation to improve the data mixture (i.e., the final training dataset and the

[4] *Data Sets: RedPajama-Data-1T*, Hugging Face, https://web.archive.org/web/20230420075601/https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T (archived Apr. 20, 2023).

[5] *Datasets: togethercomputer/RedPajama-Data-1T*, Hugging Face, https://web.archive.org/web/20240510231649/https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T (archived May 10, 2024).

proportion of each component dataset) of the model. Each experiment is also known as an "ablation" in technical parlance.

31. The first step in training the model is acquiring and curating the data that goes in to the model. This acquisition necessarily includes the copying and downloading of the data, usually onto persistent storage. This is because training an LLM or any other generative AI model is expensive and resource intensive, so retaining persistent copies of datasets avoids the need to reacquire the data when developing new models or altering the data mixtures during the development process.

32. Training an LLM is not only a function of quantity of data, but also of quality. The selection and curation of training data is therefore an important first step in training. Copyrighted books are well known among developers of AI models to be high-quality data for training LLMs.

33. After the initial copying and processing of data, the LLM copies each textual work in the training dataset and extracts protected expression from it. During what is known as *pretraining*, 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 (also known as *parameters*) that are stored within the model. These weights are entirely and uniquely derived from the protected expression in the training dataset. Once a model is pretrained, the result is a trained model known as a *base* or *foundational* model.

34. During the development process of an LLM, engineers may also conduct experiments known as ablations or "ablation studies" that test the effect of certain data on the model. This can include, for example, determining whether there is a difference in the quality of a model's output if it is trained with books or without. A dataset may be used to run such experiments but ultimately be excluded from the final training mixture of the model. Importantly, these datasets used for ablation studies may also consist of copyrighted works, including books.

35. Once an LLM has copied the textual works in the training dataset and extracted the protected expression into stored weights, an LLM is able to emit convincing simulations of natural written language in response to user prompts. Whenever an LLM generates text output in response to a

user prompt, it is performing a computation that relies on these stored weights, with the goal of imitating the protected expression ingested from the training dataset.

36. Throughout each step of the training pipeline, the same dataset may be used (i.e., copied) multiple times. Indeed, given the cost of developing an LLM, it is a ubiquitous practice to retain datasets for future use, whether that use is to pretrain other models, to perform ablations on a model, or to fine-tune an already trained base model. Each step involves making additional copies of the underlying data. The implication is that if a dataset contains unlawfully-obtained copyrighted material, each step of the training pipeline may result in an unauthorized use (i.e., infringement) of that copyrighted work.

C. Salesforce Creates a Library of Pirated Books to Develop its Large Language Models

37. Salesforce, founded in 1999, provides cloud-based services to its clients, with a particular focus on sales and e-commerce.

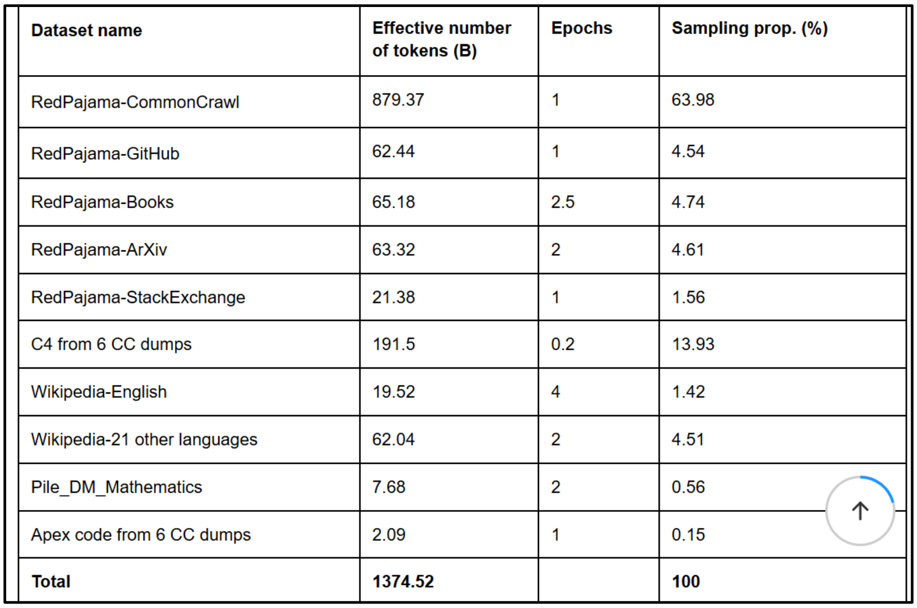
38. In March 2022, Salesforce released its CodeGen series of LLMs.[6] Salesforce admitted that "[a]ll variants of CODEGEN are firstly pre-trained on the Pile."[7] Within The Pile is the Books3 dataset containing Plaintiffs and Class members' works. Accordingly, Salesforce downloaded and stored a copy of The Pile containing Books3 before March 2022 and used it to develop and train its CodeGen series of LLMs.

39. With this library of nearly 200,000 pirated books at its disposal, Salesforce decided to use Books3 for the development and training of its next language model.

[6] Erik Nijkamp et al., *CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis* (2023), https://arxiv.org/pdf/2203.13474.pdf (on file with Joseph Saveri Law Firm, LLP).

[7] *Id.*

40. In June 2023, Salesforce released the XGen series of LLMs in a blog titled, “Long Sequence Modeling with XGen: A 7B LLM Trained on 8K Input Sequence Length” (“XGen Blog”).[8] This series consists of the XGen-7B-4K-base, XGen-7B-8K-base, and XGen-7B-{4K, 8K}-inst models. In its announcement, Salesforce provided a chart detailing the models’ pre-training data, which admits that “RedPajama-Books” was one of the training datasets:

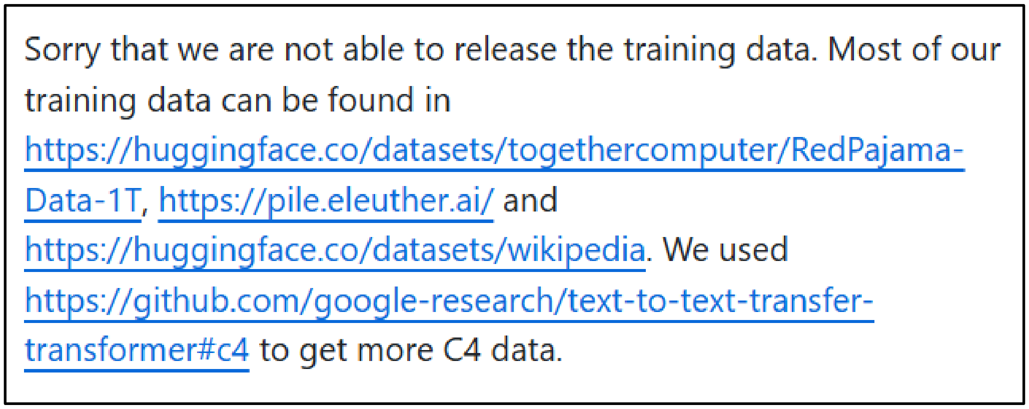


Dataset name	Effective number of tokens (B)	Epochs	Sampling prop. (%)
RedPajama-CommonCrawl	879.37	1	63.98
RedPajama-GitHub	62.44	1	4.54
RedPajama-Books	65.18	2.5	4.74
RedPajama-ArXiv	63.32	2	4.61
RedPajama-StackExchange	21.38	1	1.56
C4 from 6 CC dumps	191.5	0.2	13.93
Wikipedia-English	19.52	4	1.42
Wikipedia-21 other languages	62.04	2	4.51
Pile_DM_Mathematics	7.68	2	0.56
Apex code from 6 CC dumps	2.09	1	0.15
Total	**1374.52**		**100**

41. The RedPajama-Books dataset is a copy of the Books3 dataset that contains Plaintiffs and Class members’ copyrighted works.

[8] Erik Nijkamp et al., *Long Sequence Modeling with XGen: A 7B LLM Trained on 8K Input Sequence Length*, Salesforce AI Research Blog, https://web.archive.org/web/20230628183232/https://blog.salesforceairesearch.com/xgen/ (archived June 28, 2023).

42. At the same time that Salesforce announced the release of the XGen models, it uploaded the models to GitHub, a website where companies can provide access to open-source models and users can ask questions to their creators. On the day of the models' release, a user posed a question to Salesforce on the XGen GitHub page: "Hi, could you please release the training data too, to enable further research into the model behavior?"[9] One day later, a user by the name "tianxie-9" responded:[10]



Sorry that we are not able to release the training data. Most of our training data can be found in https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T, https://pile.eleuther.ai/ and https://huggingface.co/datasets/wikipedia. We used https://github.com/google-research/text-to-text-transfer-transformer#c4 to get more C4 data.

43. Tian Xie was a then-Salesforce employee listed as an author of the XGen Blog. In his response, he provided a link to the RedPajama dataset hosted on Hugging Face, and a link to EleutherAI's website that contains a link to download the Pile from The-Eye. Mr. Xie therefore admits that the XGen models were trained on both The Pile and RedPajama.

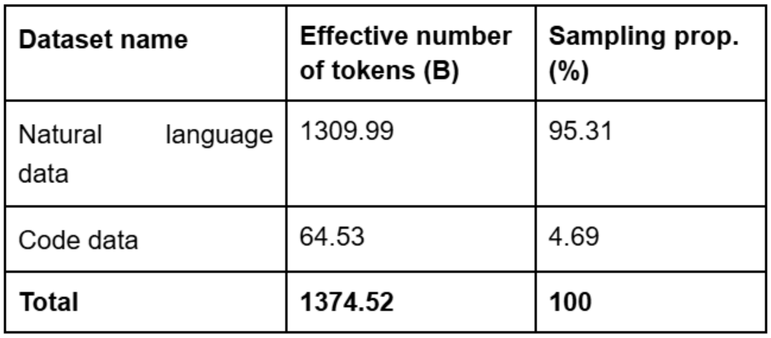
44. Because Salesforce acquired these datasets before October 2023, Salesforce copied, stored, and used versions of The Pile and RedPajama containing Books3 to develop and train its XGen models.

45. Salesforce's transparency into the training data used for its XGen models was short-lived. Two months later, in or around September 2023, it deleted from its website the chart listing the

[9] *Salesforce/xGen Release of Training Data #3*, GitHub (June 28, 2023), https://github.com/salesforce/xgen/issues/3 (on file with Joseph Saveri Law Firm, LLP).

[10] *Id.*

training data for the models. All references to "RedPajama-Books" were removed. Instead, a new chart simply listed the models as trained on "natural language data."[11]



Dataset name	Effective number of tokens (B)	Sampling prop. (%)
Natural language data	1309.99	95.31
Code data	64.53	4.69
Total	**1374.52**	**100**

46. And on September 7, 2023, Salesforce published a paper titled, "XGen-7B Technical Report."[12] The report describes the training data for the XGen models as "[n]atural language data [that] is a mixture of publicly available data." It contains no reference to the RedPajama dataset being used to develop the XGen models.

47. Then, in a December 2023 blog titled, "Developing the New XGen: Salesforce's Foundational Large Language Models," Salesforce claimed the models were trained on a "legally compliant dataset that was derived from diverse knowledge domains."[13] Per Salesforce, this training dataset comprised a "vast volume of high quality data," and was a "massive and diverse dataset, drawing from public sources."[14] In collaboration with Salesforce legal and ethics experts, Salesforce cleaned the dataset "at a tremendous scale" in order to "manage copyright issues."[15] Salesforce no longer admitted to using RedPajama to develop its models.

[11] Shafiq Joty, *Long Sequence Modeling with XGen: A 7B LLM Trained on 8K Input Sequence Length*, Salesforce Blog (June 28, 2023), https://www.salesforce.com/blog/xgen/ (on file with Joseph Saveri Law Firm, LLP).

[12] Erik Nijkamp et al., *XGen-7B Technical Report* (Sept. 7, 2023), https://arxiv.org/pdf/2309.03450.pdf (on file with Joseph Saveri Law Firm, LLP).

[13] Shafiq Joty, *Developing the New XGen: Salesforce's Foundational Large Language Models*, Salesforce Blog (Dec. 12, 2023), https://engineering.salesforce.com/developing-the-new-xgen-salesforces-foundational-large-language-models/ (on file with Joseph Saveri Law Firm, LLP).

[14] *Id.*

[15] *Id.*

48. On October 21, 2024, Salesforce released the xGen-Sales model.[16] This "finely tuned model" "was created by refining [Salesforce's] most advanced large language models with human-in-the-loop reinforcement learning, using vast datasets from diverse sources."[17] Accordingly, xGen-Sales is a fine-tuned model powered by the XGen series of LLMs that infringe on Plaintiffs and Class members' works. Upon information and belief, the xGen-Sales model was developed and trained on the RedPajama and The Pile datasets. Salesforce then commercialized these models by incorporating them into its Agentforce AI platform for enterprise customers.[18]

49. On May 2, 2025, Salesforce released the xGen-Small LLM—part of the XGen family of models.[19] The training data comprised "a wide spectrum of publicly available sources" and "natural language content."[20] Upon information and belief, Salesforce developed and trained the xGen-Small model with the RedPajama and The Pile datasets. Given that Salesforce has not publicly disclosed other datasets or sources which may have been used to train the xGen-Small model, other sources of pirated copyrighted works may have also been used to develop the xGen-Small model.

50. Salesforces infringed on Plaintiffs and Class members' copyrighted works on a massive scale. Salesforce downloaded these books from two datasets—The Pile and RedPajama—without authorization from or compensation to their authors. Salesforce then continued copying and storing the datasets, and used them to develop and train the XGen series of LLMs.

51. Upon information and belief, Salesforce used Plaintiffs and Class members' copyrighted works to train other models, including non-public models, whether sourced from RedPajama, The Pile or other pirated sources.

[16] Denise Pérez et al., *xGen-Sales: Next-Gen AI for Sales Cloud*, Salesforce Blog (Oct. 21, 2024), https://www.salesforce.com/blog/xgen-sales/ (on file with Joseph Saveri Law Firm, LLP).

[17] *Id.*

[18] *Id.*

[19] Erik Nijkamp et al., *xGen-small: Enterprise-Ready Small Language Models*, Salesforce Blog (May 2, 2025), https://www.salesforce.com/blog/xgen-small-enterprise-ready-small-language-models/ (on file with Joseph Saveri Law Firm, LLP).

[20] *Id.*

52. And, upon information and belief, the "high quality" and "natural language data" available from public sources that Salesforce used to develop its large language models is not limited to the Books3 dataset, but also potentially includes other known publicly available datasets containing unauthorized copies of copyrighted books.

D. Salesforce's Piracy And Use of Copyrighted Works Is Not Fair Use.

53. Salesforce's copying, downloading, storage, and use of Plaintiffs and Class members' works was not fair use. Salesforce could have—but chose not to—lawfully obtain the Infringed Works. 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). Salesforce downloaded hundreds of thousands of unauthorized copies from the Books3 dataset, which includes materials sourced from shadow libraries known to indiscriminately hoover up countless copyrighted works, for its centralized commercial database, all with complete disregard to the rights of copyright holders. Salesforce built its pirated library containing Plaintiffs' and Class members' copyrighted works as a substitute for authorized copies of those works. In doing so, Salesforce displaced or diluted the market for Plaintiffs' and Class members' works. Plaintiff's works were copied and maintained for future commercial purposes, including to develop its XGen series of LLMs. Pirating otherwise purchasable works is copyright infringement.

54. Any subsequent use of Salesforce's pirated library, such as training an LLM, 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.

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

56. The exploitation of Plaintiffs' Infringed Works was not indirect. It was and is a direct download and scraping of the entirety of Plaintiffs' Infringed Works into a centralized database with no transformation of form. Plaintiffs' Infringed Works were copied and maintained for future commercial purposes. Defendant's copying was not transitory as Plaintiffs' Infringed Works were not immediately destroyed, but were retained to train multiple of Salesforce's AI models, generating billions of dollars for Salesforce. Salesforce'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.

57. Salesforce then used these pirated copyrighted works to train its XGen series of LLMs. But this use of Plaintiffs' Infringed Works to train the XGen models is also not transformative. LLM outputs are the result of statistical outputs based on the training corpora. Broadly, LLMs operate by probabilistically predicting the next "token." Tokens are units of language which can be words, combinations of letters, or even a punctuation or space. Salesforce's xGen models are no different. Recent studies demonstrate that the LLMs created by the leading AI developers are capable of regurgitating substantial portions of the training data for those models, even with the implementation of safeguards designed to prevent the regurgitation of training data.[21] Upon information and belief, given that even the most advanced LLMs can recreate nearly the entirety of copyrighted books and that there is no evidence to suggest Salesforce's LLMs are as advanced, Salesforce's XGen models are capable of regurgitating substantial portions of their training data.

58. Even if the court applies the statutory fair use factors to Salesforce's unmitigated copying through downloading datasets including Books3, and its use of those datasets to train its LLMs, the application of the factors weighs 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.

[21] Ahmed Ahmed et. al., *Extracting Books from Production Language Models* (Jan. 6, 2026), https://arxiv.org/pdf/2601.02671 (on file with Joseph Saveri Law Firm, LLP). This study analyzed the output of LLMs created by Anthropic, Google, OpenAI, and xAI—models which use the same or similar architecture to the xGen models.

59. **Factor one.** The purpose and character of Salesforce's piracy is not a close call, as its use is strictly commercial and 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. Salesforce'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 Salesforce's market cap. Further, there is nothing transformative about pirating entire verbatim copies of Plaintiffs' Infringed Works to create an unauthorized commercial database rather than sourcing free market alternatives. Piracy is not transformative and is not fair use. Neither is Salesforce's training of its xGen models using copyrighted books, since models with architecture similar to the xGen models are known to regurgitate substantial portions of their training data.

60. **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. Here, Plaintiffs' 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 Salesforce chose not to purchase or license this copyrighted material. The nature of Plaintiffs' Infringed Works ensures Plaintiffs the greatest copyright protection and narrows Defendant's argument for fair use.

61. **Factor three.** The third factor asks whether a qualitatively and/or quantitatively substantial portion of Plaintiffs' Infringed Works was stolen. The entirety of Plaintiffs' Infringed Works was taken, reproduced verbatim, and stored for Salesforce'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 extracted. Plaintiffs' 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.

62. **Factor four.** The harm to Plaintiffs' 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 Salesforce's actions displaced or

diluted the market for Plaintiffs' and the Class members' Infringed Works. 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 Plaintiffs' market. At every turn, Salesforce's conduct injures Plaintiffs. Salesforce injures Plaintiffs' market when downloading a pirated library, when assembling this centralized database for ongoing commercial use, and when training its various LLMs.

63. Salesforce's unauthorized Books3 and pirated library downloads displaced Plaintiffs' 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 versions of their works. Moreover, while merely pointing towards a theoretical licensing market is insufficient, Salesforce has demonstrated that there is a thriving licensing market for Plaintiffs' Infringed Works as Salesforce has paid millions to copyright holders whose work it chose not to infringe upon. Salesforce's individual actions have injured Plaintiffs' market and widespread adoption of Salesforce's tactics will only exacerbate this harm.

64. The purpose of copyright, since the beginning of this country, is to incentivize original creativity. Salesforce's displacement of the market for the works of Plaintiffs and the Class severely disincentivizes their future creativity. Therefore, considering copyright's historic purpose, Salesforce's market displacement requires an adjudication that Salesforce's conduct is not fair use.

65. Even though Salesforce was internally using copyrighted books it obtained without authorization, compensation, or consent, its executives made statements ostensibly siding with the copyright holders.

66. For example, in a January 2024 interview, Salesforce CEO Marc Benioff acknowledged that artificial intelligence companies "ripped off" training data to build their models, and that "all the training data has been stolen."[22] Benioff also stated, "There's a pretty great company to be built on a standardized set of training data that lets all these companies play a fair . . . game and let the content

[22] *Salesforce's Benioff on Building Trust in AI*, YouTube (uploaded by Bloomberg Live, Jan. 16, 2024), https://www.youtube.com/watch?v=JSlniwSmBuI.

creators . . . get paid fairly for their work. And I think that bridge has not yet been crossed. And that's a mistake by the AI companies. Very easy to do."[23]

67. Benioff is right—technology companies like Benioff's own Salesforce that use the intellectual property of copyright holders like Plaintiffs and Class members should fairly compensate them. That is why Plaintiffs bring this lawsuit.

CLASS ALLEGATIONS

68. The "**Class Period**" as defined in this Complaint begins on at least October 15, 2022 and runs through the present. Because Plaintiffs do not yet know when the unlawful conduct alleged herein began, but, on information and belief, allege that the conduct likely began earlier than October 15, 2022, Plaintiffs reserve the right to amend the Class Period to comport with the facts and evidence uncovered during further investigation or through discovery.

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

> **All legal or beneficial owners of copyrighted works registered with the United States Copyrighted Office that Salesforce downloaded, copied, stored, or otherwise used to develop any of Salesforce's large language models during the Class Period.**

70. This Class definition excludes:

a. Defendant named herein;

b. any of the Defendant's co-conspirators;

c. any of Defendant's parent companies, subsidiaries, and affiliates;

d. any of Defendant's officers, directors, management, employees, subsidiaries, affiliates, or agents;

e. all governmental entities; and

f. the judges and chambers staff in this case, as well as any members of their immediate families.

[23] *Id.*

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

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

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

74. **Commonality** and **predominance**. Numerous questions of law or fact common to each Class member arise from Defendant's conduct and predominate over any questions affecting the members of the Class individually:

a. Whether Defendant violated the copyrights of Plaintiffs and the Class when it obtained copies of Plaintiffs' Infringed Works, stored the Infringed Works, copied the Infringed Works, and used the Infringed Works to develop Defendant's LLMs;

b. Whether any affirmative defense excuses Defendant's conduct; and

c. Whether any statutes of limitation limits the potential for recovery for Plaintiffs and the Class.

75. **Other class considerations**. Defendant has acted on grounds generally applicable to the Class. This class action is superior to alternatives, if any, for the fair and efficient adjudication of this controversy. Prosecuting the claims pleaded herein as a class action will eliminate the possibility of repetitive litigation. There will be no material difficulty in the management of this action as a class action. The prosecution of separate actions by individual Class members would create the risk of inconsistent or varying adjudications, establishing incompatible standards of conduct for Defendant.

CAUSE OF ACTION

COUNT I

Direct Copyright Infringement (17 U.S.C. § 501) against Defendant

76. Plaintiffs incorporate by reference the preceding factual allegations.

77. As the owners of the registered copyrights in the Infringed Works, Plaintiffs hold the exclusive rights to those books under 17 U.S.C. § 106.

78. Salesforce downloaded, ingested, or otherwise acquired copies of the RedPajama and The Pile datasets containing Books3, which includes the Infringed Works. The initial downloading constitutes the first unauthorized copying of Plaintiffs' works by Salesforce in the training pipeline.

79. Salesforce stored copies of RedPajama and The Pile in its internal servers.

80. To develop the XGen-7B-4K-base, XGen-7B-8K-base, and XGen-7B-{4K, 8K}-inst language models, Salesforce copied the RedPajama and the Pile datasets containing Books3 to develop these models and incorporate the datasets into the models' training dataset. Salesforce made multiple copies of these datasets containing Books3 during the development of the XGen series of LLMs.

81. Upon information and belief, Salesforce copied the RedPajama and the Pile datasets containing Books3 to develop the xGen-Sales and xGen-Small models and incorporate the datasets into the models' training dataset.

82. Plaintiffs and the Class members never authorized Defendant to make copies of their Infringed Works, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works). All those rights belong exclusively to Plaintiffs and the Class members under the U.S. Copyright Act.

83. By copying, storing, processing, reproducing, and using the datasets containing copies of Plaintiffs' Infringed Works, Defendant has directly infringed Plaintiffs' exclusive rights in their copyrighted works.

84. By copying, storing, processing, and reproducing the XGen models trained on Plaintiffs' Infringed Works, Salesforce has directly infringed Plaintiffs' exclusive rights in their copyrighted works.

85. Defendant repeatedly copied, stored, and used the Infringed Works without Plaintiffs' permission. Defendant made these copies without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act.

86. Because the training mixture for the xGen-Sales, xGen-Small and other nonpublic Salesforce models has not been made public, other unauthorized repositories of pirated copies of Plaintiffs' and Class members' copyrighted works may have been used to develop Salesforce's xGen-Sales, xGen-Small, and other models.

87. Defendant's infringing conduct alleged herein was and continues to be willful and carried out with full knowledge of Plaintiffs' rights in the copyrighted works. As a direct result of their conduct, Defendant has wrongfully profited from copyrighted works that they do not own.

88. By and through the actions alleged above, Defendant has infringed and will continue to infringe Plaintiffs' copyrights.

89. Plaintiffs have been and will continue to be injured by Defendant's acts of direct copyright infringement. Salesforce's infringement has directly resulted in and is directly causing ongoing harm, including, but not limited to (1) loss of value to Plaintiffs and Class members' works due to Salesforce's mass infringement that has replaced and depressed the overall market for the Plaintiffs' and Class members' respective commercial markets, and (2) lost licensing revenue that Plaintiffs and Class members would have received had Salesforce properly obtained authorization, consent, or license to acquire and use their works.

90. Plaintiffs are entitled to statutory damages, actual damages, restitution of profits, and all appropriate legal and equitable relief.

DEMAND FOR JUDGMENT

Wherefore, Plaintiffs request that the Court enter judgment on their behalf and on behalf of the Class defined herein, by ordering:

a) This action may proceed as a class action, with Plaintiffs serving as Class Representatives, and with Plaintiffs' counsel as Class Counsel.

b) Judgment in favor of Plaintiffs and the Class and against Defendant.

c) An award of statutory and other damages under 17 U.S.C. § 504 for violations of the copyrights of Plaintiffs and the Class by Defendant.

d) Reasonable attorneys' fees and reimbursement of costs under 17 U.S.C. § 505 or otherwise.

e) A declaration that such infringement is willful.

f) Destruction or other reasonable disposition of all copies Defendant made or used in violation of the exclusive rights of Plaintiffs and the Class, under 17 U.S.C. § 503(b).

g) Pre- and post-judgment interest on the damages awarded to Plaintiffs and the Class, and that such interest be awarded at the highest legal rate from and after the date this FACC is first served on Defendant.

h) Defendant is responsible financially for the costs and expenses of a Court-approved notice program through post and media designed to give immediate notification to the Class.

i) Further relief for Plaintiffs and the Class as may be appropriate.

JURY TRIAL DEMANDED

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

Dated: January 20, 2026

By: *_/s/ Joseph R. Saveri_*

Joseph R. Saveri (SBN 130064)
Christopher K.L. Young (SBN 318371)
William W. Castillo Guardado (SBN 294159)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
cyoung@saverilawfirm.com
wcastillo@saverilawfirm.com

Lesley E. Weaver (SBN 191305)
Anne K. Davis (SBN 267090)
Joshua D. Samra (SBN 313050)
STRANCH, JENNINGS & GARVEY, PLLC
1111 Broadway, Suite 03-130
Oakland, CA 94607
Tel. (341) 217-0550
Email: lweaver@stranchlaw.com
adavis@stranchlaw.com
jsamra@stranchlaw.com

Karin B. Swope (*pro hac vice*)
Thomas E. Loeser, (SBN 202724)
Andrew Fuller (*pro hac vice* to be filed)
Jacob M. Alhadeff (*pro hac vice* to be filed)
COTCHETT, PITRE & MCCARTHY, LLP
1809 7th Ave., Ste. 1610
Seattle, WA 98101
Tel: (206) 802-1272
Fax: (206) 299-4184
Email: kswope@cpmlegal.com
tloeser@cpmlegal.com
afuller@cpmlegal.com
jalhadeff@cpmlegal.com

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
COTCHETT, PITRE & MCCARTHY, LLP
840 Malcom Road
Burlingame, CA 94010
Tel: 650-697-6000
Fax: 650-697-0577
Email: jcotchett@cpmlegal.com
bdanitz@cpmlegal.com
gjung@cpmlegal.com
cyuen@cpmlegal.com

Joseph I. Marchese (*pro hac vice* to be filed)
Julian C. Diamond (*pro hac vice* to be filed)
Caroline C. Donovan (*pro hac vice* to be filed)
BURSOR & FISHER, P.A.
1330 Avenue of the Americas, 32nd Floor
New York, NY 10019
Tel: (646) 837-7150
Fax: (212) 989-9163
Email: jmarchese@bursor.com
jdiamond@bursor.com
cdonovan@bursor.com

Attorneys for Individual and Representative Plaintiffs E. Molly Tanzer, Jennifer Gilmore, Tasha Alexander, Jon McGoran, and Art Kleiner