IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MONTANA
BUTTE DIVISION

DARIUS H. JAMES, individually
and on behalf of a class of similarly
situated individuals,

Plaintiff,

v.

SNOWFLAKE INC., a Delaware
corporation,

Defendant.

Case No._CV-25-108-BU-BMM

CLASS ACTION COMPLAINT

JURY TRIAL DEMANDED

CLASS ACTION COMPLAINT

Plaintiff Darius H. James ("Plaintiff"), on behalf of himself and all others similarly situated, bring this class action complaint ("Complaint") against Defendant Snowflake Inc. ("Snowflake" or "Defendant").

SUMMARY OF THE CASE

1. Artificial intelligence ("AI") refers to software engineered to mimic human-like reasoning and inference through algorithmic processes, typically leveraging statistical methods.

2. Large language models ("LLMs") are AI software programs designed to reply to user prompts with natural-sounding text outputs. Snowflake's "Arctic" models are a family of LLMs trained, created, and then released by Snowflake for enterprise AI use, including general text generation.

3. While the traditional coding process involves human coders inputting explicit instructions, an LLM is instead trained by processing vast quantities of text from diverse sources (a "training dataset"), learning statistical patterns and associations within that data, and encoding

those abstract representations into a vast array of numerical values known as parameters. The goal is to enable the model to learn general language patterns, grammar, factual knowledge, and contextual relationships. This results in a versatile base model that can understand and generate human-like text.

4. Creating a high-quality training dataset necessarily involves copying an enormous quantity of textual works. Each book or other text in the dataset must be downloaded, copied, stored, and processed (often multiple times) in order to be tokenized, filtered, deduplicated, and ingested in a large-scale pretraining and training process.

5. The training dataset used by Snowflake to train its LLMs used public domain, licensed, and, crucially, unlicensed copyrighted materials.

6. Plaintiff and Class members are authors. They own registered copyrights in certain books (the "Infringed Works") that were included in the training dataset that Snowflake pirated and copied from the internet to train its Arctic LLMs. Plaintiff and Class members never authorized Snowflake to download, copy, store, and use their copyrighted works as training materials. Snowflake copied, and thus infringed on, these copyrighted works multiple times to train its Arctic LLMs.

7. Through the acts described in further detail below, Defendant has infringed on Plaintiff's copyrighted works and it continues to do so by continuing to store, copy, use, and process the training datasets containing copies of Plaintiff's and the putative Class's Infringed Works.

JURISDICTION AND VENUE

8. 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).

9. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. §§ 1391(b)(2) and (c)(2) as Snowflake is headquartered in Bozeman, Montana and thus is headquartered in this District. 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 the illegal scheme and conspiracy 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.

PARTIES

10. Plaintiff Darius H. James is an author and performance artist who resides in Connecticut. He is the author of two American publications, *Negrophobia* and *That's Blaxploitation!: Roots of the Baadassss 'Tude,* and two bilingual German publications, *Voodoo Stew* and *Froggie Chocolates' Christmas Eve*.

11. A list of Plaintiff's registered copyrights owned are attached hereto as Exhibit A.

12. Defendant Snowflake is a Delaware corporation with its principal place of business at 106 E. Babcock Suite 3A, Bozeman, MT 59715.

13. The unlawful acts alleged against the Defendant in this 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 business or affairs.

14. 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.

FACTUAL ALLEGATIONS

15. LLMs are trained by ingesting massive training corpora consisting of extremely large volumes of text – often millions or billions of lines. Constructing these corpora is accomplished by acquiring and digitally copying copyrighted works and storing those copies, oftentimes in multiple locations and formats, to support preprocessing, deduplication, tokenization, and training. Sometimes these digital copies are acquired legally, other times, as alleged here, copyrighted works are illegally pirated from the internet.

16. During pre-training, the LLM processes each textual work in the training dataset to learn statistical patterns and associations within it. The LLM adjusts its parameters through optimization techniques to get progressively better at predicting sequences in the data, capturing general linguistic structures rather than specific expressions. The results of this learning process are encoded in a large set of numbers called parameters stored within the model. These parameters are derived from the entire pre-training dataset.

17. The RedPajama dataset is a training dataset assembled and published by Together Computer, Inc. that contained within it a deduplicated copy of the Books3 dataset. Books3 was described in a paper by EluetherAI called "*The Pile: An 800GB Dataset of Diverse Text for Language Modeling*" as follows:

> 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. [1]

18. The RedPajama dataset, including Books3, was available for download from Hugging Face (a website dedicated to "a mission to democratize good machine learning") as a standalone dataset and was downloaded tens of thousands of times.[2][3]

19. In addition, while the Hugging Face website currently states that Books3 was eventually removed from RedPajama due to "reported" copyright infringement, an archived version of that Hugging Face webpage explicitly disclosed that Books3 was included in the RedPajama corpus of materials available for download and training.[4]

20. Plaintiff's copyrighted books are among the works in the RedPajama dataset.

21. Snowflake is best known as a provider of cloud-based data warehousing and analytics services. In recent years, Snowflake has aggressively expanded into generative AI and LLMs in an effort to maintain competitiveness and increase revenue.[5]

22. In 2023 Snowflake began assembling and curating a large-scale pretraining corpora to assist in developing its Arctic family of LLMs. It downloaded, assembled, copied, and stored such large-scale datasets, including the RedPajama dataset.

[1] https://arxiv.org/abs/2101.00027 (last accessed Nov. 18, 2025)

[2] https://huggingface.co/huggingface (last accessed Nov. 18, 2025)

[3] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T (last accessed Nov. 18, 2025)

[4] https://web.archive.org/web/20230920185843/https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T (last accessed Nov. 18, 2025)

[5] https://www.snowflake.com/en/blog/generative-ai-llms-summit-2023/ (last accessed Nov. 18, 2025)

23. At its Snowflake Summit 2023, Defendant unveiled its new wave of AI products including multiple LLM products and services which were trained on the RedPajama dataset that included Plaintiff's and members of the Class's Infringed Works.

24. On April 24, 2024, Snowflake publicly launched its "Arctic" LLM as a cornerstone of its AI offerings. Third-party technical summaries and Snowflake-focused commentary describe Arctic as being trained over the course of three months using key public datasets including RedPajama.[6]

25. Thus, in order to train Arctic and its related family of LLMs, Snowflake downloaded, copied, stored, and used the RedPajama dataset that contained Books3 and Plaintiff's Infringed Works. Snowflake also repeatedly downloaded, copied, and processed those works during the preprocessing and pretraining of the models.

26. Snowflake retained copies of those pretraining datasets that contained copies of the Infringed Works on its servers and continues to store and use them in further training for new versions of its Arctic LLMs and related models at a minimum through retaining the model parameters of its originally trained model.

27. Thus, Defendant directly infringed on Plaintiff's and Class Members' copyrighted works on a massive scale. Snowflake downloaded and copied these books and the Infringed Works as contained in the RedPajama dataset without authorization from, or after providing compensation to, their authors. Snowflake then continued copying and storing the datasets and used them to develop and train its Arctic LLMs and other related models.

[6] https://medium.com/%40m_chak/meet-snowflake-arctic-llm-perfect-snowflake-and-sql-copilot-50b69e7bf279 (last accessed Nov. 18, 2025)

CLASS ALLEGATIONS

28. The "Class Period" as defined in this Complaint begins on at least November 25, 2022, and runs through the present. Because Plaintiff does not yet know when the unlawful conduct alleged herein began, but believes, on information and belief, that the conduct likely began prior to November 25, 2022, Plaintiff reserves the right to amend the Class Period to comport with the facts and evidence uncovered during further investigation or through discovery.

29. Plaintiff seeks certification of the following Class pursuant to Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3):

> All persons or entities domiciled in the United States that own a United States copyright in any work that was copied, stored, or used as training data by Defendant during the Class Period.

30. Plaintiff will fairly and adequately represent and protect the interests of the other members of the Class. Plaintiff has retained counsel with substantial experience in prosecuting complex litigation and class actions. Plaintiff and his counsel are committed to vigorously prosecuting this action on behalf of the other members of the Class, and have the financial resources to do so. Neither Plaintiff nor his counsel have any interest adverse to those of the other members of the Class.

31. Absent a class action, most members of the Class would find the cost of litigating their claims to be prohibitive and would have no effective remedy. The class treatment of common questions of law and fact is also superior to multiple individual actions or piecemeal litigation in that it conserves the resources of the courts and the litigants and promotes consistency and efficiency of adjudication.

32. Defendant has acted and failed to act on grounds generally applicable to Plaintiff and the other members of the Class, requiring the Court's imposition of uniform relief to ensure

compatible standards of conduct toward the members of the Class, and making injunctive or corresponding declaratory relief appropriate for the Class as a whole.

33. The factual and legal bases of Defendant's liability to Plaintiff and to the other members of the Class are the same, resulting in injury to Plaintiff and to all of the other members of the Class. Plaintiff and the other members of the Class have all suffered harm and damages as a result of Defendant's unlawful and wrongful conduct.

34. There are many questions of law and fact common to the claims of Plaintiff and the other members of the Class, and those questions predominate over any questions that may affect individual members of the Class. Common questions for the Class include, but are not limited to, the following.

a. Whether Defendant violated the copyrights of Plaintiff and the Class by obtaining and creating copies of Plaintiff's Infringed Works with the intent to use the Infringed Works for commercial benefit;

b. Whether Defendant did use the Infringed Works of Plaintiff and the Class for commercial benefit;

c. Whether Defendant violated the copyrights of Plaintiff and the Class by using illicitly obtained copies of Plaintiff's Infringed Works to train Defendant's AI models; and

d. Whether Defendant caused further infringement of the Infringed Works by distributing its AI models under an open license.

FIRST CAUSE OF ACTION
Direct Copyright Infringement,
(17 U.S.C. § 501)
(On Behalf of Plaintiff and the Class)

35. Plaintiff repeats the allegations contained in the foregoing paragraphs as if fully set forth herein.

36. Plaintiff, as the owner of registered copyrights, holds the exclusive rights to those books under 17 U.S.C. § 106.

37. In order to supply enough data for pre-training and training of the Snowflake Arctic LLMs and other related models, Defendant downloaded, copied, and stored copies of the RedPajama dataset which included the Books3 dataset as a subset. Thus, the RedPajama dataset includes Plaintiff's copyrighted works. Defendant made multiple copies of the dataset (and thus Plaintiff and the Class Members' copyrighted works) for pre-training and training its models.

38. Neither Plaintiff nor Class Members authorized Defendant to make copies of, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works) of their copyrighted works. The U.S. Copyright Act bestows all the aforementioned rights only on Plaintiff and the Class Members.

39. By copying, storing, processing, reproducing, and using the datasets containing copies of Plaintiff's Infringed Works, Defendant has directly infringed on Plaintiff's exclusive rights in his copyrighted works.

40. Defendant repeatedly copied, stored, and used the Infringed Works without Plaintiff's and members of the Class's permission in violation of their exclusive rights under the Copyright Act.

41. By and through the actions alleged above, Defendant has infringed and will continue to infringe on Plaintiff's copyrights.

42. Plaintiff has been injured by Defendant's acts of direct copyright infringement. Plaintiff is entitled to statutory damages, actual damages, restitution of profits, and all appropriate legal and equitable relief.

PRAYER FOR RELIEF

WHEREFORE, Plaintiff, individually and on behalf of all others similarly situated, seeks the following against Defendant:

a. An order certifying the Class, naming Plaintiff as Class Representative, and naming Plaintiff's attorneys as Class Counsel to represent the Class;

b. An order declaring that Defendant's conduct violates 17 U.S.C. § 501;

c. An award of statutory and other damages under 17 U.S.C. § 504 for violations of the copyrights of Plaintiff 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 Plaintiff and the Class, under 17 U.S.C. § 503(b);

g. Pre- and post-judgment interest on the damages awarded to Plaintiff and the Class, and that such interest be awarded at the highest legal rate from and after the date this class action complaint is first served on Defendant; and

h. Further relief for Plaintiff and the Class as the Court deems may be appropriate.

JURY TRIAL DEMAND

Plaintiff demands a trial by jury on all causes of action and issues so triable.

(signature on following page)

DATED: November 21, 2025. Respectfully submitted,

/s/ *John Heenan*
John Heenan
HEENAN & COOK PLLC
1631 Zimmerman Trail
Billings, MT 59102
Telephone: (406) 839-9091
john@lawmontana.com

Myles McGuire (pro hac vice forthcoming)
David L. Gerbie (*pro hac vice* forthcoming)
Jordan R. Frysinger (*pro hac vice* forthcoming)
MCGUIRE LAW, P.C.
55 W. Wacker Drive, 9th Floor
Chicago, IL 60601
Telephone: (312) 893-7002
mmcguire@mcgpc.com
dgerbie@mcgpc.com
jfrysinger@mcgpc.com

Counsel for Plaintiff and the Proposed Class