Eugene Y. Turin (SB # 342413)
McGuire Law, P.C.
1089 Willowcreek Road, Suite 200
San Diego, CA 92131
Tel: (312) 893-7002 Ex. 3
Fax: 312-275-7895
eturin@mcgpc.com

Counsel for Plaintiff and the Putative Class Members

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

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

Plaintiff,

v.

TOGETHER COMPUTER, INC., a Delaware corporation,

Defendant.

Case No.

COMPLAINT

CLASS ACTION

DEMAND FOR JURY TRIAL

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

OVERVIEW

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 ("LLM") are AI software programs designed to reply to user prompts with natural-sounding text outputs.

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. These parameters emerge from the collective structure of the entire training dataset, which often blends public domain, licensed, and, crucially, unlicensed copyrighted materials. When an LLM generates text in response to a prompt, it performs probabilistic computations over these parameters to produce coherent, contextually relevant output that generalizes from the learned patterns.

4. Plaintiff and Class members are authors. They own registered copyrights in certain books (the "Infringed Works") that were included in the open training dataset, RedPajama, that Together AI assembled and published for the purpose of training LLMs, and ultimately driving such LLM development to its own development platform. Plaintiff and Class members never authorized Together AI to download, copy, store, publish or use their copyrighted works to create the RedPajama dataset.

5. By the acts described in further detail below, Together AI infringed on

Plaintiff's copyrighted works, and it may continue to do so if it continues to download, copy, store, publish and use the RedPajama dataset which contains copies of Plaintiff's and the putative Class's Infringed Works.

JURISDICTION AND VENUE

6. 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).

7. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Together AI is headquartered in this district. Together AI copied Plaintiff's and Class members' Infringed Works and assembled the RedPajama dataset. 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.

8. 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

9. Plaintiff Darius H. James is an author and performance artists 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*.

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

A.

11. Defendant Together AI is a Delaware corporation with its principal place of business at 251 Rhode Island Street, Suite 205, San Francisco, CA 94103.

12. 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. 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.

13. 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, or for Defendant with respect to the acts, violations, and common course of conduct alleged herein.

FACTUAL ALLEGATIONS

14. Together AI is a company that contributes open-source research, models, and datasets to other AI-focused companies and sells access to a cloud platform to help other entities train, fine-tune, and deploy generative AI models.[1]

15. In an effort to drive more AI startups to its cloud platform, TogetherAI orchestrated and assembled the RedPajama dataset, a dataset comprised of a mixture of publicly available works as well as copyrighted works.

16. The RedPajama dataset assembled and published by TogetherAI 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:

[1] https://www.together.ai/about

> 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]

17. Before October 2023, Books3 was available for download from Hugging Face (a website dedicated to "a mission to democratize good machine learning") as a standalone dataset.[3] 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."[4]

18. Plaintiff's copyrighted books are among the works in the Books3 dataset.

19. Books3 was also downloaded and copied by Defendant and utilized to create and publish its RedPajama dataset in April of 2023.

20. The RedPajama dataset contained a subset called "Books" or "RedPajama-Books" that was actually a copy of the Books3 dataset. Specifically, the RedPajama dataset "is a publicly available, fully open, best-effort reproduction of the training data. . . used to train the first iteration of LLaMA family of models."[5] This LLaMA training dataset included the Books3 section of The Pile (a broader publicly available dataset which contained Books3).[6]

21. Thus, RedPajama training datasets, including the RedPajama-Books subset, contained copyrighted works, including books written and copyrighted by

---

[2] https://arxiv.org/abs/2101.00027

[3] https://huggingface.co/huggingface

[4] https://web.archive.org/web/20231127101818/https:/huggingface.co/datasets/the_pile_books3

[5] https://arxiv.org/html/2411.12372v1

[6] https://arxiv.org/pdf/2302.13971

Plaintiff and Class members.

22. Together AI was chiefly responsible for the creation of the RedPajama dataset, acknowledging itself as collaborator in its blog post and announcing the project publicly on the X platform on April 17, 2023.[7][8] Importantly, the RedPajama dataset is hosted on Defendant's specific GitHub organization page ("togethercomputer")[9] and its affiliation with the RedPajama dataset page is evidenced by the license notes disclosing "Copyright 2023 Together Computer."[10]

23. Defendant's blog post announcing the creation of the RedPajama dataset also provides a link to download the dataset on its Hugging Face "togethercomputer" profile page.[11]

24. While the Hugging Face website currently states that Books3 was removed from RedPajama due to "reported" copyright infringement, an archived version of that Hugging Face webpage from April 2023 explicitly disclosed that Books3 was included in the RedPajama corpus of materials.[12][13]

25. On this same archived webpage, it also states that "we [Together AI] use simhash to remove near duplicates" from Books3, meaning that Together AI downloaded, copied, deduplicated, repackaged and re-published Books3 within its RedPajama dataset online for others to download and train their LLMs.[14]

26. Accordingly, through its creation and publication of the RedPajama dataset, and other data subsets including RedPajama-Books, Defendant has itself

---

[7] https://www.together.ai/blog/redpajama

[8] https://x.com/togethercompute/status/1647917989264519174

[9] https://github.com/togethercomputer/RedPajama-Data

[10] *Id*.

[11] https://www.together.ai/blog/redpajama

[12] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T

[13] https://web.archive.org/web/20230417120911/https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T

[14] *Id*.

downloaded, copied, stored, and used Books3 and the copyrighted works of Plaintiff and the members of the Class.

27. Furthermore, Defendant has actively promoted the RedPajama dataset and incentivized third parties to utilize, download, and copy it. For example, in May 2023, Together AI announced a monetary incentive, hosted by Chai Research, encouraging third parties to "create the best chatbot by fine-tuning RedPajama-INCITE-3B, with a $1M prize for the winner!"[15]

28. Between April 2023 and October 2023, Together AI's promotions incentivized over 190,000 downloads of the RedPajama dataset within the Together AI community and third parties.[16]

29. Thus, from time that Together AI made RedPajama available on the Hugging Face and GitHub websites in April of 2023, to the time that Books3 was removed from the dataset, Defendant made available for download, and promoted the downloading and copying of a vast repository of texts including the copyrighted works of Plaintiff and the members of the Class.

CLASS ALLEGATIONS

30. The "Class Period" as defined in this Complaint begins on at least November 5, 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 5, 2022, Plaintiff reserves the right to amend the Class Period to comport with the facts and evidence uncovered during further investigation or through discovery.

[15] https://www.together.ai/blog/redpajama-3b-updates#:~:text=,tune%20for%20longer%20context

[16] https://www.together.ai/blog/redpajama-data-v2#:~:text=Over%20the%20last%20half%20a,released%20specifically%20for%20LLM%20training

31. 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 downloaded, copied, stored, used, or contained within the RedPajama training datasets during the Class Period.

32. 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 their 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 their counsel have any interest adverse to those of the other members of the Class.

33. Absent a class action, most members of the Class would find the cost of litigation 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.

34. 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.

35. 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.

36. 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 downloading, copying, storing, and using the Infringed Works to create its training datasets;

b. Whether Defendant knew of and materially contributed to the downloading, storage, retention, and copying of the Infringed Works by third parties;

c. Whether Plaintiff and members of the Class are entitled to actual and/or statutory damages for the aforementioned violations; and

d. Whether Defendant caused further infringement of the Infringed Works by distributing the pre-training dataset under an open license.

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

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

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

39. In order to create a comprehensive training dataset known as RedPajama, Together AI downloaded and copied Books3. Books3 includes the Plaintiff's and the members of the Class's Infringed Works. Together AI copied, downloaded, used, stored, and republished many copies of the Books3 dataset and thus Plaintiff's and the Class members' Infringed Works.

40. Neither Plaintiff nor Class members authorized Together AI to make

copies, publicly display copies, or distribute copies of their Infringed Works. The U.S. Copyright Act bestows all the aforementioned rights only on the Plaintiff and the Class members.

41. By downloading, copying, storing, processing, reproducing, and using the datasets, like Books3, containing copies of Plaintiff's Infringed Works, Defendant has directly infringed upon Plaintiff's exclusive rights in his copyrighted works.

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

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

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

45. 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.

SECOND CAUSE OF ACTION
Contributory Copyright Infringement
17 U.S.C. § 501 *et seq.*
(On behalf of Plaintiff and the Class)

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

47. Unknown third parties directly infringed on the Plaintiff's and Class members' rights by downloading, copying, using, storing and retaining the Infringing Works. Such use is not fair use and is infringing. Each and every one of these infringements is facilitated, encouraged, and made possible by Defendant.

48. Defendant had actual and/or red-flag knowledge that the pirated library that it downloaded, copied, stored, duplicated and republished through the RedPajama dataset contained copyrighted works.

49. Defendant had actual and/or red-flag knowledge that third parties were downloading and copying its RedPajama dataset that it created using the copyrighted works of Plaintiff and the members of the Class.

50. Defendant materially contributed to unknown third party's infringement by: 1) creating the RedPajama dataset which contained copyrighted works; 2) providing its RedPajama dataset to the public for unregulated copying and downloading on both its own specific GitHub and Hugging Face pages; and 3) promoting and encouraging unknown third parties to copy and download its RedPajama dataset.

51. Defendant's contribution was substantial and necessary to the scope and persistence of the infringement. Plaintiff and the Class members suffered damages, including statutory damages and profits attributable to the infringement.

PRAYER FOR RELIEF

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

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

b. For 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 awards 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;

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.

DATED: November 6, 2025 Respectfully submitted,

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

By: /s/ *Eugene Y. Turin*
Plaintiff's Attorney

Eugene Y. Turin (SB # 342413)
David L. Gerbie (*pro hac vice* forthcoming)
Jordan R. Frysinger (*pro hac vice* forthcoming)
McGuire Law, P.C.
1089 Willowcreek Road, Suite 200
San Diego, CA 92131
Tel: (312) 893-7002 Ex. 3
Fax: 312-275-7895
eturin@mcgpc.com
dgerbie@mcgpc.com
jfrysinger@mcgpc.com

Counsel for Plaintiff and the Putative Class

EXHIBIT A

Registration Number / Date: TX0003372564 / 1992-08-05

Title: *Negrophobia: an urban parable: a novel* / Darius James.

Copyright Claimant:

Darius James

USCO Catalog Link:

https://publicrecords.copyright.gov/detailed-record/voyager_14732620

Registration Number / Date: TX0004193939 / 1996-01-03

Title: *That's blaxploitation!: roots of the baadassss 'tude* / Darius James.

Copyright Claimant: Darius James

USCO Catalog Link:

https://publicrecords.copyright.gov/detailed-record/voyager_15167875