UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC. COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: All Matters	No. 25-md-3143 (SHS) (OTW)

BRIEF OF ENGINE ADVOCACY, ET AL. AS AMICI CURIAE IN SUPPORT OF DEFENDANTS

Marc Hearron
HEARRON PLLC
305 Broadway, Suite 700
New York, New York 10007
(212) 897-5847
marc@hearronlaw.com

Counsel for Amici Curiae Engine Advocacy, et al.

TABLE OF CONTENTS

TABLE OF AUTHORITIES ii

INTEREST OF AMICI CURIAE 1

INTRODUCTION AND SUMMARY OF ARGUMENT 2

ARGUMENT 4

I. Using Copyrighted Content to Train AI Systems Is a Lawful Fair Use 4

A. Startups rely heavily on AI models trained on enormous swaths of potentially copyrighted materials 4

B. Startups' use of AI to create their products shows how using potentially copyrighted works to train AI models is a fair use 7

II. Whether the Largest AI Players Are Liable for Copyright Infringement Will Affect the Pace of Innovation by Startups, Which Play a Vital Role in the AI Ecosystem 10

A. If licensing were required, the staggering costs would lock out many startups from the AI ecosystem 10

B. Requiring licenses for training data would curtail startup innovation in other ways 14

C. Licensing the works used to train a foundation model is not realistically possible 17

D. These results would be sharply at odds with the purpose of copyright law 19

CONCLUSION 19

CERTIFICATE OF COMPLIANCE 21

TABLE OF AUTHORITIES

Cases

Andy Warhol Found. for Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023) 8

Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) 9, 19

Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) 9

Bartz v. Anthropic PBC, 787 F. Supp. 3d 1007 (N.D. Cal. 2025) 9, 12

Bartz v. Anthropic PBC, No. 4:24-cv-05417-AMO (N.D. Cal. July 20, 2026), ECF No. 680 12

Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994) 7

Cox Commc'ns, Inc. v. Sony Music Ent., 146 S. Ct. 959 (2026) 16

Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) 11, 19

Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021) 8, 9, 15, 19

Kadrey v. Meta Platforms, Inc., 788 F. Supp. 3d 1026 (N.D. Cal. 2025) 9

Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003) 9

Stewart v. Abend, 495 U.S. 207 (1990) 19

Constitutional Provisions

U.S. Const. art. I, § 8, cl. 8 19

Statutes

17 U.S.C. § 102(a) 11

17 U.S.C. § 107 9

17 U.S.C. § 504(c) 16

Other Authorities

Rebecca Beyer, *License to Stream*, The Record, B.U. Sch. of L. (May 26, 2022) 18

Michael Dorf, *Music Performing Rights Organizations Exacerbate Inequality Among Artists*, Michael Dorf Uncorked (June 4, 2025) 18

Engine, *AI & Copyright: The Infeasibility of Licensing and the Impact on Startups* (Oct. 29, 2026) 11, 12, 14, 15, 18

Engine, *#StartupsEverywhere: Austin, Texas* (July 17, 2026) 5, 7

Engine, *#StartupsEverywhere: Durham, N.C.* (Feb. 14, 2025) 5, 6, 15

Engine, *#StartupsEverywhere: New York, N.Y.* (Nov. 21, 2025) 5, 6

Engine, *#StartupsEverywhere: Raleigh, N.C.* (Apr. 10, 2025) 5, 6, 7

Engine, *Mapping the Future: Charting the AI Ecosystem & a Policy Blueprint for Startup Success* (May 2025) 14

Engine, *The State of the Startup Ecosystem* (Apr. 2021) 14

FAQs on the Authors Guild's Positions and Advocacy Around Generative AI, Authors Guild (last visited Oct. 29, 2026) 12, 17

Cameron Miller et al., *The Economic Importance of Fair Use for the Development of Generative Artificial Intelligence* (Data Catalyst Inst. June 2025) 17

Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Nw. U. L. Rev. 1607 (2009) 16

Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45 (2017) 16

World Bank, *World Development Indicators: GDP* (2025) 12

INTEREST OF AMICI CURIAE[1]

Engine Advocacy ("Engine") is a nonprofit technology policy, research, and advocacy organization dedicated to bridging the gap between startups and policymakers. Engine works with government officials and a community of thousands of high-technology, growth-oriented startups across the nation to support innovation and entrepreneurship through research, policy analysis, and advocacy. Engine has long advocated for a legal framework that allows startups to build, train, and use artificial intelligence ("AI"). Engine's advocacy around AI includes addressing the issue currently before this Court; in comments filed with the U.S. Patent and Trademark Office and the U.S. Copyright Office, Engine has explained that using copyrighted material to train AI models is lawful and that a contrary rule would put AI development out of reach for all but the largest companies.

Developers Alliance is an advocate for software developers, the companies they lead, and the industries that depend on them. Developers Alliance works with policymakers to share the needs and perspectives of developers. Developers Alliance advocates for a legal and regulatory environment that supports developers in creating and advancing new technologies that benefit the public and business alike.

Lazy Moose Co. is a startup that deploys Xaidus, a software platform that helps teens accomplish goals and gives parents insight into their progress and influences. The company uses proprietary technologies and leverages large-language models, including those made by OpenAI, to provide real-time analysis while protecting user privacy. The

[1] No party's counsel authored this brief in whole or in part, and no party, party's counsel, or person other than amici, their members, or their counsel contributed money intended to fund the preparation or submission of this brief.

company relies on frontier large-language models' capabilities available to them at a low cost to speed user onboarding and ensure product quality. A ruling that undermines the availability, cost, or quality of frontier AI models would harm Lazy Moose Co.'s competitiveness, because without access to the advanced capabilities of frontier models, the company projects a 300% increase in user onboarding time and a 75% decrease in consumer satisfaction with their systems.

INTRODUCTION AND SUMMARY OF ARGUMENT

Although this copyright dispute focuses on the use of potentially copyrighted material by some of the largest AI technology companies, the AI ecosystem is much larger and more diverse than its largest players, and the outcome of this case will have far-reaching implications for much smaller businesses. By volume, startups make up the vast majority of the AI ecosystem, but their perspectives can be drowned out by attention on the largest players.

Startups are using AI in innovative ways in communities and industries across the country. They are doing so by taking existing general-purpose foundation models, fine tuning them for specific tasks, and building businesses that offer niche, industry-specific AI-based products. These products may answer the phones for restaurants so that staff can cook and tend to guests, answer residents' 311 calls for municipalities, help auto-parts sellers make sense of their sales data, or help manufacturers find forever chemicals buried in their supply chains.

Startups' use of foundation models demonstrates that using copyrighted works to train an AI model is a fair use, not copyright infringement. The purpose of training a foundation model is not to reproduce the works it learns; it is to build a general-purpose

tool that others can adapt to new ends. The ends startups have chosen bear no resemblance to the purposes of the works in any training dataset. Take, for example, a startup's AI product that answers restaurant phones and takes dinner orders; that has nothing to do with the purposes of the underlying works used to train the AI models. No one calls a restaurant instead of buying a book or subscribing to a newspaper. Training is transformative, and it takes nothing from the market for the works on which models learn.

If the Court were to hold otherwise, licensing would become an insurmountable barrier to AI development among startups. Licensing the works needed to train a single large-language model would cost on the order of $10 trillion, and developers passing those costs downstream in whole would need to charge each commercial user roughly $1.4 million—about the size of an average startup's entire first funding round, and an impossible sum for a seed-stage company. But that is not all. Large-language model developers that could not afford these exorbitant costs—especially makers of open-source models, which startups depend on—would need to reduce the amount of data used to train their models, resulting in models of severely degraded quality to the point of being unsuitable for use. Even if foundation models could continue to be used, a startup that fine-tunes a foundation model to develop its product would need its own licenses for the data needed for the fine-tuning process—another barrier to innovation. And statutory damages of up to $150,000 per work, applied to datasets containing billions of works, could deter investors from funding AI startups at all.

This, of course, assumes that all the copyrighted works needed for training data could be licensed, but that is an impossible task. Billions of potentially copyrightable works are needed to train a single large-language model. The rightsholders for all those works cannot realistically be identified, let alone negotiated with one by one.

If using copyrighted content to train AI models constitutes copyright infringement and is not a fair use, the AI ecosystem will become unaffordable for most startups, and innovation by startups will be stifled. That result is at odds with the purpose of copyright, which is not to reward authorship for its own sake but to promote progress for the benefit of the public. The Court should hold that using copyrighted works to train AI models is a fair use.

ARGUMENT

I. Using Copyrighted Content to Train AI Systems Is a Lawful Fair Use

Startups' use of foundation AI models, including OpenAI's, to build their own tools and products demonstrates that using copied material to train AI models is a fair use. The use is transformative, creating brand new products that bear essentially no resemblance to the original content and are used for wholly different purposes. Such use also does not affect the market for the copyrighted materials because these new tools are in no way a substitute for the original works.

A. Startups rely heavily on AI models trained on enormous swaths of potentially copyrighted materials

Startups today are implementing AI in innovative ways to develop industry-specific products. Ninety percent of the startups that joined Engine's network in the past two years use AI in their main product. Many of these startups are using AI to build the product that the startup offers to its own customers or to build tools that the startup uses to run its business.

Take one example from Engine's startup network: Maple. Based in New York City, Maple built an AI-based phone-answering service for restaurants. Its product integrates into a restaurant's existing orders and reservations systems and uses AI to take customer

orders, book and modify reservations, and manage delivery requests. Its founder, who comes from a family of Chinese restauranteurs, recognized that restaurants often miss calls and lose orders because staff are too busy. Maple boosts restaurant productivity by handling the restaurant's phone lines, freeing up staff to focus on cooking and in-person hospitality. Maple is also using AI to expand into other restaurant workflows, such as intelligent waitlist management and taking supplier calls. Engine, *#StartupsEverywhere: New York, N.Y.* (Nov. 21, 2025)[2] [hereinafter *Maple Profile*].

Engine's network includes other examples of AI startups. CivicReach.AI answers 311 calls for municipalities and can provide information to city residents, fill out forms, and schedule city services. Engine, *#StartupsEverywhere: Raleigh, N.C.* (Apr. 10, 2025)[3] [hereinafter *CivicReach Profile*]. Tromml's AI-based product helps companies in the automotive aftermarket industry make sense of their sales data. Engine, *#StartupsEverywhere: Durham, N.C.* (Feb. 14, 2025)[4] [hereinafter *Tromml Profile*]. And EcoPulse uses AI in its product to help large manufacturers manage complex supply chain risks by identifying the presence of PFAS, or "forever chemicals," across tens of thousands of product components. Engine, *#StartupsEverywhere: Austin, Texas* (July 17, 2026)[5] [hereinafter *EcoPulse Profile*].

Startups like these are not building their own foundation AI models from scratch. Instead, they are leveraging existing foundation models—accessing open-source models or licensing closed models—to build their own unique products. Maple, for example, did not build its restaurant phone-answering tool by first developing a general AI system that

[2] https://www.engine.is/news/startupseverywhere-new-york-ny-maple.
[3] https://www.engine.is/news/startupseverywhere-raleigh-nc-civicreach-ai.
[4] https://www.engine.is/news/startupseverywhere-durham-nc-tromml.
[5] https://www.engine.is/news/startupseverywhere-austin-tx-ecopulse.

can recognize and parse human speech. Rather, Maple started with an already-developed open-source AI model for the core of its product, and it leveraged large-language models to train and fine-tune the core model to adapt it to the various tasks that Maple's product would need to perform. *Maple Profile*, *supra*.

This process is known as fine-tuning: adapting a pretrained AI model to perform specific tasks. Fine-tuning leverages the general knowledge already acquired by the foundation model during its initial training, and it specializes the model for the startup's specific needs. Fine-tuning an existing model accelerates innovation by drastically reducing costs and reducing barriers to entry. The startup can circumvent the high costs and complexities of building a brand-new AI model, which would require vast computational resources, access to enormous quantities of data, and significant time—resources that most startups do not have. Fine-tuning also enables startups to build better products by enabling the developers to focus their limited resources on their true innovation rather than on building foundational technology. As Tromml's CEO explained, "[n]o one is building an LLM from scratch when open-source models provide a foundation to build upon. The real innovation comes from how these models are fine-tuned and applied to specific industries." *Tromml Profile*, *supra*.

A startup may use multiple foundation models, because each one may excel at a particular task, and combine them to build the startup's own tool. As mentioned, Maple uses a smaller, specialized AI model for the core of its product, and it uses large-language models to fine-tune the core model. *Maple Profile*, *supra*. CivicReach uses several large-language models, including OpenAI's, as the base layer for its product; its own proprietary models interact with the LLMs so that each customer can choose the model it is most comfortable with. *CivicReach Profile*, *supra*.

Fine-tuning relies on foundation models' adaptability to a wide variety of tasks. That adaptability is the result of foundation models' having been built using enormous amounts of diverse training data—most of which is potentially copyrightable. The diversity of this data is key: without it, foundation models would not be adaptable and could not be fine-tuned for specific purposes. The same foundation model that one startup adapts to answer a restaurant's phones might be adapted by another startup to identify forever chemicals in supply chains and adapted by a third startup to analyze auto-parts sales.

The fine-tuning process may require access to training data that is potentially copyrighted. For example, CivicReach builds each municipality a tailored data system so that its product draws insights from city-council minutes, websites, social media, and press releases. *CivicReach Profile*, *supra*. EcoPulse's product that identifies PFAS in its customers' component materials analyzes supply chain data as well as detailed chemical information and research findings. *EcoPulse Profile*, *supra*. Such materials may be publicly available, but public availability does not mean that a work is free of copyright.

B. Startups' use of AI to create their products shows how using potentially copyrighted works to train AI models is a fair use

The examples discussed above demonstrate that the use of potentially copyrighted material to train or fine-tune an AI model is a fair use because it is transformative and because it does not affect the market for the original work.

Transformative. The first fair-use factor asks whether the use "merely supersedes the objects of the original creation or instead adds something new, with a further purpose or different character." *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994) (quotation marks, brackets, and citation omitted). This factor "relates to the

problem of substitution"; the "use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work is more likely to substitute for, or supplant, the work." *Andy Warhol Found. for Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 528 (2023) (quotation marks and brackets omitted).

Using copies of copyrighted works to train or fine-tune AI models is "[e]xceedingly [t]ransformative." Statement of Interest of U.S. at 10, ECF No. 1682. AI models do not merely replicate the content they were trained on; they generate new outputs for a different purpose and with a different character. Through training, a model learns relationships and patterns among words, pixels, and tones, and once training is finished, it produces new outputs informed by those relationships. *See id.* (explaining that a large-language model "uses the copyrighted work not to duplicate the work's expressive content, but as part of a process to learn and act on statistical patterns in written text, including vocabulary, syntax, and knowledge").

The startups described above prove the point. Their products may be built on large-language models that were trained on copious numbers of copyrighted newspaper articles and novels. The purpose of a newspaper article is to inform its readers of the news, and the purpose of a novel is to entertain. Maple's product has nothing to do with these purposes; its purpose is to answer a restaurant's telephone calls, place customer orders, and book table reservations. Likewise, an AI agent that answers a resident's call to a city hotline, flags a manufacturer's PFAS risk, or analyzes an auto parts seller's sales data serves a purpose that bears no resemblance to the purpose of any book or article in a foundation model's training data.

Courts have recognized that much less transformational use of copyrighted materials meets the first fair-use factor. In *Google*, the Supreme Court held that Google's

copying of Sun's Java code was transformative, even though it was copied "in part for the same reason that Sun created" the code, because Google's use sought "to create new products," namely "a highly creative and innovative" platform that programmers could readily use for a smartphone environment. *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 30 (2021). Such copying was "consistent with that creative 'progress' that is the basic constitutional objective of copyright itself." *Id.* Making thumbnail copies of images to help index and improve access to content on the Internet is also transformative. *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818–19 (9th Cir. 2003). And making a digital copy of a copyrighted book so that it is easier for people to search and identify books "involves a highly transformative purpose." *Authors Guild v. Google, Inc.*, 804 F.3d 202, 216 (2d Cir. 2015). That the copying encompassed entire works did not change the analysis: "Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original." *Id.* at 221. It is no surprise, then, that other courts addressing the training of large language models have likewise found that use highly transformative. *See Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1021 (N.D. Cal. 2025); *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1044 (N.D. Cal. 2025).

No market effect. Such use also does not hamper "the potential market for or value of the copyrighted work." 17 U.S.C. § 107. This factor "is concerned with only one type of economic injury to a copyright holder: the harm that results because the secondary use serves as a substitute for the original work." *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 99 (2d Cir. 2014). "[A]ny economic 'harm' caused by transformative uses does

not count because such uses, by definition, do not serve as substitutes for the original work." *Id.*

The use of copyrighted content to train, test, or fine-tune an AI model does not harm the creator's ability to sell or license the original content; an AI system does not sell, license, or make publicly available the underlying original content. Startups are not using AI foundation models to create copies of the original, copyrighted works to compete with those works in the market. They are creating tools that generate completely new outputs that in no way substitute for the original works. Customers call a restaurant in Maple's network to order food or make a reservation; no one calls one of those restaurants instead of buying a novel. And no city resident calls a CivicReach-powered 311 system instead of subscribing to a newspaper.

II. Whether the Largest AI Players Are Liable for Copyright Infringement Will Affect the Pace of Innovation by Startups, Which Play a Vital Role in the AI Ecosystem

A. If licensing were required, the staggering costs would lock out many startups from the AI ecosystem

Contrary to the News Plaintiffs' suggestion that copyright "licensing fees are not barriers to entry into the AI industry in any meaningful sense," News Pls.' Resp. to Stmt. of Interest of U.S. at 10, ECF No. 2071, requiring large AI companies to pay copyright licenses will throttle the pace of AI innovation among startups. There is no realistic way for developers of foundation models to license all the copyright-eligible works in their training data, but even if it could be done, the costs would be so exorbitant that developers would have to pass them on to their commercial users. That would put large-language models out of reach for most startups and mean that startups would be unable to afford to participate in the AI ecosystem going forward. *See* Engine, *AI & Copyright: The*

Infeasibility of Licensing and the Impact on Startups 2–3 (Oct. 29, 2026)[6] [hereinafter *Engine Report*].

The number of potentially copyrighted works used to train a foundation model is vast. A large-language model learns by ingesting enormous amounts of written language and mapping the relationships among words so that it can predict which word is most likely to follow another in a string of text. To do that for every word a user might supply, a model must ingest an unfathomable quantity of language, and developers accomplish that by ingesting not just lengthy works like books and plays but also short-form content like forum posts, encyclopedia entries, subtitles, and captions on photos and videos. *See id.* at 1.

The diversity of the training data required to build a foundation model means that enormous swaths of that data are likely to consist of copyrighted material. Almost any expressive content on the Internet is potentially copyrightable, from news articles and songs to meaningless social-media posts about brushing one's teeth. *See* 17 U.S.C. § 102(a) (extending protection to "original works of authorship fixed in any tangible medium of expression"). While underlying facts, data, and ideas are not eligible for copyright protection, the works containing that underlying training material may be. *See Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 349–50 (1991). So the comment at the bottom of a blog post is as eligible for copyright protection as a novel published by a major publisher. The trillions of terabytes of expression on the internet are thus largely copyright-eligible, and the universe of rightsholders is correspondingly vast.

[6] https://www.engine.is/news/category/ai-copyright-the-infeasibility-of-licensing-and-the-impact-on-startups.

In a mandatory-licensing regime, a developer of a large-language model would need to license more than three billion works just to train a single model. *See Engine Report*, *supra*, at 2–3.[7] At $3,000 per work—the approximate per-work recovery in the recently approved settlement in the *Bartz* litigation[8]—licensing the pretraining data for one model would cost more than $10 trillion. Far from "barely a footnote," *contra* News Pls.' Resp. at 10, $10 trillion is equivalent to the combined annual gross domestic product of roughly 160 of the world's economies, *see Engine Report*, *supra*, at 3 (citing World Bank, *World Development Indicators: GDP* (2025)[9]). And the figure may be understated, because some rightsholders assert that training is not a one-time event and that licensing fees should continue for as long as the model remains in use. *See FAQs on the Authors Guild's Positions and Advocacy Around Generative AI*, Authors Guild[10] [hereinafter *Authors Guild FAQs*].

News Plaintiffs point to the size and capital expenditures of "*current* AI model developers"—around $1 trillion in valuation and combined capital expenditures of $1.1 trillion over a three-and-a-half-year period—to argue that licensing fees are not

---

[7] This estimate assumes that 30 trillion pretraining tokens (the units of text that a model processes and produces) are used to train a single model—which is based on the number of tokens that were used to develop open models in 2025. *Id.* at 2. Thirty trillion tokens is roughly 22.5 trillion words, which, at an average of 1,000 words per work, is some 22.5 billion individual works. *Id.* at 2–3. Assuming conservatively that fifteen percent of those works are commercially prepared and in need of a license, 3.375 billion works would need to be licensed to train one model. *Id.* at 3.

[8] Order Granting Final Approval of Class Action Settlement at 4, *Bartz v. Anthropic PBC*, No. 4:24-cv-05417-AMO (N.D. Cal. July 20, 2026), ECF No. 680. That settlement resolved claims that copies of books had been downloaded from pirate sources, not claims about the use of books to train a model; the court held that the training itself was fair use. *See Bartz*, 787 F. Supp. 3d at 1021.

[9] https://databank.worldbank.org/id/8cc72e5#.

[10] https://authorsguild.org/advocacy/artificial-intelligence/faq/ (last visited Oct. 29, 2026).

meaningful barriers to entry into the AI industry. News Pls.' Resp. at 10–11 (emphasis added). But as shown above, $1 trillion is just a fraction of what it would cost to obtain licenses to all the data that would be needed to train a single large-language model. While it may be possible for the largest AI developers to reach licensing agreements with the largest rightsholders—likely at a still-exorbitant cost that no startup or new market entrant could afford—that would be nowhere close to enough data. The largest rightsholders' works would be a mere drop in the bucket of the total works needed to train a large-language model. If Plaintiffs' position in this case were adopted, it would incentive anyone who has authored something original and posted it online to demand copyright royalties—leading to the nearly unbounded potential licensing costs described above and erecting insurmountable barriers to entry.

The alternative would be for AI developers to severely restrict the amount of data they use to train models. But that would lead to severe degradation in quality because the amount of data a model is trained on is related to how well it performs. These worse-performing foundation models would harm the startup innovation that would otherwise be built on them. Startups need to be confident that the models used in their products are sufficiently capable before deploying them—e.g., the developer of a phone-answering product needs to be sure its product can competently understand a customer's order or a user's request. This degradation in quality would also be a barrier to entry for startups and would quash innovation.

Even assuming the largest model developers and their largest customers could afford steep increases in cost, foundation models would become prohibitively expensive for startups. Model developers would need to defray their extraordinary licensing costs by passing them on to their commercial users through higher prices. Spread across the

roughly 7 million U.S. businesses that use AI, the per-commercial-user licensing cost for a single model would be approximately $1.4 million. *See Engine Report*, *supra*, at 3. This amount is roughly the size of an average startup's entire initial funding round. *See* Engine, *The State of the Startup Ecosystem* 5 (Apr. 2021).[11] The average seed-stage startup has about $55,000 per month to cover all of its expenses, including salaries, rent, equipment, research and development, and customer acquisition. *Id.* at 17. A startup operating on that kind of budget cannot absorb a significant increase in the cost of the foundation models on which its product depends.

B. Requiring licenses for training data would curtail startup innovation in other ways

If using copyrighted material for training data is infringement and not a fair use, it would effectively put an end to open-source foundation models, which startups rely on heavily. An open-source AI model is one that its developer makes publicly available for anyone to download, use, modify, and build upon, typically by releasing the model's weights—the numerical parameters learned during training that determine how the model behaves—often along with the model's code and documentation of the data used to train it. *See* Engine, *Mapping the Future: Charting the AI Ecosystem & a Policy Blueprint for Startup Success* 8, 14–15 (May 2025).[12] Open-source models are crucial for startups: because the functionality of an AI model lies primarily in the configuration of its weights, a publicly available pretrained model lets a startup sidestep the data and compute required to train a model from scratch and focus its limited resources on its own

[11] https://www.engine.is/news/engine-releases-report-on-the-health-of-the-startup-ecosystem.

[12] https://www.engine.is/news/category/mapping-the-future-charting-the-ai-ecosystem-a-policy-blueprint-for-startup-success.

innovation. Open-source developers, however, will have little reason to release a model for free and with documentation revealing how it was trained: doing so would expose those developers to infringement liability or licensing costs while leaving them with no way of recouping those costs. That would be a serious blow to startups. Tromml's CEO explained that "[w]ithout open-source models and tools when we started, we would have been years behind on deployment," and that "[t]hese tools are crucial for startups, enabling us to adapt, innovate, and bring solutions to market faster." *Tromml Profile*, *supra*.

Potential infringement liability would also make it infeasible for startups to create their own tools by fine-tuning foundation models. The $10 trillion estimate discussed above is only for training a foundation model. A startup that fine-tunes a foundation model to develop its own unique product often uses its own dataset of copyright-eligible works for the fine-tuning process; those works would have to be licensed separately, at an additional expense. *See Engine Report*, *supra*, at 3. Startups have incredibly limited time and resources, and it would be impractical or impossible for them to obtain a license every time potentially copyrighted material is included in the training data used to develop their product. To license one copyrighted work, a company has to (1) figure out whether the work is protected by copyright, (2) find out who owns it, (3) track that owner down, and (4) negotiate a price, if possible. As the Supreme Court has recognized, copyright protection "can impose special costs, such as the cost of contacting owners to obtain reproduction permission." *Google*, 593 U.S. at 17. Those costs add up quickly. One analysis of a hypothetical licensing regime for Google Books estimated that clearing the rights to the roughly 8.4 million books with some potential copyright constraint would cost over a billion dollars even if clearance cost only $200 per book, and could exceed $10

billion, before a single royalty was paid. *See* Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Nw. U. L. Rev. 1607, 1659–61 (2009). Those figures are from 2009, and today's training datasets are measured in billions of works, orders of magnitude more than that example.

The alternative to licensing would be to use the content without permission and risk infringement litigation. Because of the sheer number of potentially copyrighted works involved, the resulting exposure could be crippling. A finding of infringement carries statutory damages of $750 to $30,000 per work, and up to $150,000 per work for willful infringement. 17 U.S.C. § 504(c)(1)–(2). A copyright owner may elect those damages "instead of actual damages and profits," *id.* § 504(c)(1), so they are available even if the infringing use had little to no effect on the market for the underlying work. The stakes are not hypothetical: in one case involving 10,017 works, a jury awarded $1 billion in statutory damages. *Cox Commc'ns, Inc. v. Sony Music Ent.*, 146 S. Ct. 959, 966 (2026) (describing verdict before reversing on liability). As one scholar observed before current model architectures were widely used and before the current wave of litigation, "[b]ecause machine learning datasets can contain hundreds of thousands or millions of works, an award of statutory damages could cripple even a powerful company." Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 80 (2017). Current models are orders of magnitude larger, making the risk of statutory damages existential.

The specter of statutory damages at this scale is so enormous as to disincentivize investment in AI startups, were the Court to rule for Plaintiffs. With the scale of licenses that startups would need for their training data, uncertainty about the feasibility of obtaining those licenses and about how much they would cost could alone be enough to

cripple interest from investors. And if the amount of potential copyright infringement liability is more than the entire value of a small- or medium-sized company, that could easily shut off investment opportunities for startups.

C. Licensing the works used to train a foundation model is not realistically possible

Even setting aside the costs of licenses, the act of obtaining licenses from all the potential rightsholders would be impossible. Even for content that is created and distributed commercially, where ownership information is comparatively well documented and centralized, a developer would have to contact and negotiate with hundreds of thousands of authors and publishers. But those are not the only rightsholders; there are also the hundreds of millions of people creating tangible expressions in myriad ways online. Identifying and locating rightsholders at that scale is not feasible at all. And a developer that managed to find them would still have to paper each agreement; by one estimate, the cost of having an attorney draft and negotiate the individual licensing agreements alone would run into the millions of dollars. *See* Cameron Miller et al., *The Economic Importance of Fair Use for the Development of Generative Artificial Intelligence* 21 (Data Catalyst Inst. June 2025).[13] Even the Authors Guild, which advocates for licensing, concedes that it "is not efficient or even practicable for AI companies to seek licenses from each individual author who owns the rights to their works." *Authors Guild FAQs*, *supra*.

Collective licensing is no answer either. Group licensing systems, like the ones that license music for radio play, work only when there is a defined group of similar works and

[13] https://ccianet.org/research/case-studies/economic-importance-of-fair-use-for-development-of-generative-artificial-intelligence/.

a defined group of users who all want the same kind of license. That is not how AI training data works: almost any expressive content on the Internet could end up as training data, so a collective license would only ever cover part of the data a developer needs, leaving the rest to be negotiated work by work. Existing collective licensing regimes also illustrate the problems that scaling them up would create. They can impede competition, they are notoriously difficult for small businesses and new market entrants to navigate, and the imbalance of power and information between collecting organizations and small firms leaves those firms facing disproportionate litigation threats without the resources to negotiate or litigate. *See generally* Rebecca Beyer, *License to Stream*, The Record, B.U. Sch. of L. (May 26, 2022);[14] Michael Dorf, *Music Performing Rights Organizations Exacerbate Inequality Among Artists*, Michael Dorf Uncorked (June 4, 2025).[15] And building a new licensing system for AI training would take years, time that startups do not have.

Even if a licensing mechanism could be built, there would be no sensible way to price it. A model's ability to predict the next word in a sequence comes from recognizing patterns across as many as trillions of instances of a word appearing alongside other words. Additional works give the model more to learn from, but no single work is necessary to train a model, and the value of any one work is both minimal and nearly impossible to measure: it varies with the length of the work, the uniqueness of its word choice, and the ways in which the model will be used. *See Engine Report*, *supra*, at 2.

[14] https://www.bu.edu/law/record/articles/2022/license-to-stream/.
[15] https://michaeldorf.substack.com/p/music-performing-rights-organizations.

D. These results would be sharply at odds with the purpose of copyright law

The predictable result of a mandatory-licensing regime would be to stifle innovation. Without access to foundation models, the startups currently using AI in innovative ways in communities and industries around the country would not be able to build their tools.

That result would be fundamentally at odds with the purpose of copyright law. "The primary objective of copyright is not to reward the labor of authors, but '[t]o promote the Progress of Science and useful Arts.'" *Feist*, 499 U.S. at 349 (quoting U.S. Const. art. I, § 8, cl. 8). The fair-use doctrine, by ensuring that "a copyright holder cannot prevent another person from making a 'fair use' of copyrighted material," avoids "stifl[ing] the very creativity which that law is designed to foster." *Google*, 593 U.S. at 18 (quoting *Stewart v. Abend*, 495 U.S. 207, 236 (1990)). Thus, "while authors are undoubtedly important intended beneficiaries of copyright, the ultimate, primary intended beneficiary is the public." *Authors Guild v. Google*, 804 F.3d at 212.

Treating AI training as infringement would make copyright "a lock limiting the future creativity of new programs," *Google*, 593 U.S. at 39, and would price nearly every startup out of being able to build its products with these innovative AI tools. That would be a real cost to scientific and technological progress and the opposite of what copyright law is meant to encourage.

CONCLUSION

For the foregoing reasons, the Court should hold that the use of copyrighted works to train AI models is a fair use and grant Defendants' motions for summary judgment on that issue.

Dated: October 29, 2026

Respectfully submitted,

/s/ Marc Hearron

Marc Hearron
HEARRON PLLC
305 Broadway, Suite 700
New York, New York 10007
(212) 897-5847
marc@hearronlaw.com

Counsel for Amici Curiae Engine Advocacy, et al.

CERTIFICATE OF COMPLIANCE

I certify that this brief complies with the Court's order of August 7, 2026, which permits a proposed amicus brief of no more than 6,000 words, ECF No. 1663, and with the formatting requirements of Local Civil Rule 7.1(b). The brief contains 5,637 words.

Dated: October 29, 2026

/s/ Marc Hearron