SENATE PRIVACY, DIGITAL TECHNOLOGIES, AND CONSUMER PROTECTION COMMITTEE
Senator Christopher Cabaldon, Chair
2025-2026 Regular Session

AB 412 (Bauer-Kahan)
Version: May 28, 2026
Hearing Date: June 8, 2026
Fiscal: No
Urgency: No
CK

SUBJECT

Generative artificial intelligence: training data: copyrighted materials

DIGEST

This bill places obligations on Generative AI (GenAI) developers to document "covered material" used to train their models and to facilitate a process for "rights owners" to determine whether their protected material was so used.

EXECUTIVE SUMMARY

The use of copyrighted works in training AI models has sparked significant legal and ethical debate. GenAI, in particular, can produce text, images, video, or audio that emulates the expressive works used to train it. Key concerns include whether training on copyrighted material without permission constitutes fair use or copyright infringement, particularly when models can sometimes reproduce portions of their training data. Content creators argue they should be compensated when their work is used to train commercial AI systems, while some worry that AI-generated content could compete with and displace original creative works. There are also questions about transparency – many users and creators are unaware of which copyrighted materials were included in training datasets. Over 50 lawsuits from authors, artists, and publishers are currently testing these issues in court, seeking to establish clearer legal boundaries around AI training practices. The debate balances innovation and the advancement of AI technology against intellectual property rights and the economic interests of content creators. This bill, co-sponsored by SAG-AFTRA and the Concept Art Association, establishes a framework to allow copyright owners to determine whether their registered works were used to train a GenAI model. It is supported by a number of creators' associations, advocacy organizations, and labor groups. The bill is opposed by a number of industry associations, including Technet and CalBroadband. Should the bill pass out of this Committee, it will next be heard in the Senate Judiciary Committee.

PROPOSED CHANGES TO THE LAW

Existing federal law:

1) Provides that Congress has the power to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries. (U.S. Const., art. I, § 8, cl. 8.)

2) Establishes the Copyright Act, which grants an owner of copyright to exclusive right to do and authorize reproduction of the work, derivative works, distribution of copies of the work, and performances or displays of the work. (17 U.S.C. § 106.)

3) Provides that an infringer of copyright is liable for either the copyright owner's actual damages and any additional profits of the infringer or statutory damages for all infringements involved in the action, with respect to any one work, for between $750 and $30,000 as the court considers just. The law provides for injunctive relief, costs and attorney's fees, and, in certain circumstances, criminalizes willful copyright infringement for commercial advantage or private financial gain. (17 U.S.C. §§ 502(a), 504(c), 505.).

4) Provides that nothing in the Copyright Act limits any rights or remedies under State laws with respect to, among other things, activities violating legal or equitable rights that are not equivalent to any of the exclusive rights specified above. (17 U.S.C. § 301(b)(3).)

Existing state law:

1) Requires, on or before January 1, 2026, a developer of a GenAI system or service that was released on or after January 1, 2022, to post on its website a high-level summary of whether the datasets used in the development of the system or service include, among other things, data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain. (Civ. Code § 3111.)

This bill:

1) Requires GenAI developers to do all of the following:
    a) Document covered materials the developer knows were used by the developer to train the GenAI model, and to make reasonable efforts to identify and document any other such covered materials and the rights owner of each material documented pursuant hereto.
    b) Make available on their websites:

i. Information sufficient to enable generation of a fingerprint, defined as an abstract representation of digital content that encodes distinctive features of the content, as provided, that is compatible with any covered materials used to train the GenAI model and is generated according to widely accepted industry standards.
ii. A mechanism allowing a rights owner to submit a request for information about the developer's use of covered materials. This mechanism must allow the rights owner to submit specified materials, including identity documentation and registration information and fingerprints for their covered material.

c) Document any requests received using the mechanism above.
d) Retain the above documentation as long as the developer uses the GenAI model commercially in California or makes it available to Californians for use, whichever is longer, plus five years.

2) Requires a developer, within 30 days of receiving a request for information from a rights owner using the above mechanism, to do the following:
   a) Assess whether covered material represented by a fingerprint provided by the owner is likely to be present in the developer's dataset, as provided.
   b) Provide the rights owner with a list of covered materials held by the owner that are known to be used in training the model, as specified, and a list of covered materials that a fingerprint assessment suggests are likely to be present in the developer's dataset.

3) Defines the relevant terms, including:
   a) "Rights owner" means the owner of a "covered material."
   b) "Developer" means a business, person, partnership, corporation, or other entity that designs, codes, produces, or substantially modifies a GenAI model and uses the GenAI model commercially in California or makes it available to Californians for use.
   c) "Covered material" means either of the following:
      i. A material registered, preregistered, or indexed with the United States Copyright Office pursuant to the federal Copyright Act of 1976.
      ii. A sound recording fixed before February 15, 1972, enforceable under the Classics Protection and Access Act.
      iii. Covered material does not include a material that is in the public domain.

4) Provides that a developer's collection, use, retention, and sharing of information from a rights owner pursuant hereto shall be reasonably necessary and proportionate to achieve the purposes for which the information was collected and processed, or for another disclosed purpose that is compatible with the

context in which the information was collected, and not further processed in a manner that is incompatible with those purposes.

5) Prohibits rights owners from submitting more than one request per calendar quarter to the same developer concerning the same GenAI model, unless the subsequent request includes material new information not available to the rights owner at the time of the prior request.

6) Authorizes a good-faith rights holder to bring a cause of action against developers in violation hereof for the greater of $1,000 or actual damages, injunctive or declaratory relief, attorney's costs and fees, and any other relief the court deems appropriate. Each day after the 30-day period described above that a developer fails to provide a rights owner with the information required constitutes a discrete violation.

7) Excludes GenAI models that are:
   a) Trained exclusively using datasets the developer makes publicly available at no cost to users.
   b) Trained exclusively using datasets a third party makes publicly available at no cost to users, as disclosed by the developer pursuant hereto.
   c) Developed and used solely by universities or government entities exclusively for noncommercial academic or governmental research.
   d) Not trained using covered materials.
   e) Trained exclusively using covered materials for which the developer is the rights owner.
   f) Trained exclusively using covered materials the developer licensed for the disclosed purpose of training a GenAI model.
   g) Developed and used exclusively for the operation of aircraft in the national airspace.

8) Clarifies that nothing therein shall be construed to impose liability on the provider of a telecommunications service, information service, or cable service, as those terms are defined in 47 U.S.C. § 153, for content provided by another person, to the extent the provider is not acting as the developer of a GenAI model.

COMMENTS

1. Training data

Training data is the veritable secret sauce for AI systems. Ultimately, GenAI models are only as good as the data used to train them. However, there is very little transparency in what data is used to train these systems, and that lack of transparency hamstrings

efforts to address and adequately identify many of the issues being raised by GenAI's rapid development.

GenAI models "learn" from examples such as books, articles, photos, film, or music. This learning occurs within systems of interconnected numerical parameters known as "neural networks" that encode statistical patterns gleaned from data. During training, data is broken into fundamental units known as "tokens" – syllables, pixels, or musical notes, for example – that can be represented numerically. The neural network is exposed to sequences of tokens and is prompted to predict the next most likely token. If the prediction is incorrect, the error is calculated and the model adjusts the strengths of the connections between its parameters to improve its next prediction. The process continues iteratively until the neural network can reliably emulate the human-created content it was trained on. A trained neural network embedded in a GenAI system is known as its "model," and the strengths of its connections are known as its "model weights."[1]

Staggering quantities of data are required to train the most advanced models. For example, GPT4 – the large language model (LLM) embedded in ChatGPT 4 – is reported to have been trained on roughly 10 trillion words of text, which broke down into 13 trillion tokens.[2] Adjusting the model's 1.8 trillion parameters continuously as it was exposed to this vast corpus required trillions upon trillions of computations, which were performed by running approximately 25,000 expensive, energy-consuming microchips for nearly 100 days nonstop, at an estimated cost of $63 million.3

With the race to build bigger and better AI systems and models, a battle over data is being waged:

> The race to lead A.I. has become a desperate hunt for the digital data needed to advance the technology. To obtain that data, tech companies including OpenAI, Google and Meta have cut corners, ignored corporate policies and debated bending the law, according to an examination by The New York Times.
>
> At Meta, which owns Facebook and Instagram, managers, lawyers and engineers last year discussed buying the publishing house Simon & Schuster to procure long works, according to recordings of internal meetings obtained by The Times. They also conferred on gathering

[1] IBM, *What is generative AI?*, https://www.ibm.com/think/topics/generative-ai; IBM, *What is machine learning?*, www.ibm.com/topics/machine-learning. All internet citations are current as of June 1, 2026.
[2] Maximilian Schreiner, *GPT-4 architecture, datasets, costs and more leaked* (July 11, 2023) The Decoder, https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/; Mathina Begum, *OpenAI Releases GPT-4: A Smarter and Faster AI-Language Model with 'Human-level Performance'* (2023) Vocal Media, https://vocal.media/01/open-ai-releases-gpt-4-a-smarter-and-faster-ai-language-model-with-human-levelperformance.

> copyrighted data from across the internet, even if that meant facing lawsuits. Negotiating licenses with publishers, artists, musicians and the news industry would take too long, they said.
>
> Like OpenAI, Google transcribed YouTube videos to harvest text for its A.I. models, five people with knowledge of the company's practices said. That potentially violated the copyrights to the videos, which belong to their creators.[3]

Requiring transparency about the training data used for AI systems helps identify and mitigate biases, addressing hallucinations and other problematic outputs, and shines a light on various other issues, such as privacy and, most relevant here, copyright concerns. A team of experts from both industry and academia at the Shorenstein Center created a documentation framework for AI, which highlights the importance:

> While often categorized as technical, AI systems and their underlying data and models are sociotechnical. In other words, they combine the technical infrastructure and design with the social context in which they are designed, developed, evaluated, and deployed. Accountability for these systems and their impacts requires transparency around their design and creation and how they are intended to be used. In recent years, alongside the exponential increase in data collection and the efforts to develop increasingly powerful machine learning models, there have been notable efforts calling attention to the need for documentation to accompany datasets, models, and AI systems, and to account for the process of creating them.
>
> Documentation is worthwhile for various stakeholders. It improves the understanding of practitioners creating or building datasets, models, or AI systems, which opens up opportunities to reflect on implicit and explicit decisions, ultimately enhancing the reliability of the systems they create. For organizations, it enables knowledge transfer across silos and encourages responsible use. Further, it provides information to users and potentially affected communities that can be used to determine the appropriateness of an AI system or its underlying data or models, thus helping inform consumer choice, advocacy work, regulation development, and regulation enforcement. It also enables recourse in the event of harms caused by or inquiries into the AI system, and accountability regarding who might be held responsible for those harms.[4]

---

[3] Cade Metz, et al., *How Tech Giants Cut Corners to Harvest Data for A.I.* (April 6, 2024) The New York Times, https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

[4] Kasia Chmielinski, et al., *The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers* (May 21, 2024) Shorenstein Center on Media, Politics, and Public Policy, https://shorensteincenter.org/wp-content/uploads/2024/05/CleAR_KChmielinski_FINAL.pdf.

An article in the Harvard Data Science Review highlights the importance of transparency and barriers to achieving it without regulation:

> Knowing what is in the data sets used to train models and how they have been compiled is vitally important. Without this information, the work of developers, researchers, and ethicists to address biases or remove harmful content from the data is hampered. Information about training data is also vital to lawmakers' attempts to assess whether foundation models have ingested personal data or copyrighted material. Further downstream, the intended operators of AI systems and those impacted by their use are far more likely to trust them if they understand how they have been developed.
>
> However, in undertaking their analysis, Schaul et al. (2023) concluded that "many companies do not document the contents of their training data—even internally—for fear of finding personal information about identifiable individuals, copyrighted material and other data grabbed without consent."
>
> In public, companies have used different arguments to justify the lack of transparency around their training data. In documentation published at the launch of its GPT-4 model, OpenAI (2023) stated that it would not share detailed information about 'data set construction' and other aspects of the model's development due to "the competitive landscape and the safety implications of large-scale models." The decision not to disclose the data used to train the model was roundly criticized by a number of leading researchers (Xiang, 2023). A recent op-ed in the Guardian argued that companies are using 'speculative fears' to "stop people asking awkward questions about how this particular technological sausage has been made" (Naughton, 2023).[5]

2. <u>GenAI's impact on creative economies</u>

The introduction of GenAI is reshaping California's creative economies. Early evidence suggests major disruption may be underway, and the true impact on jobs, incomes, and long-term careers remains largely unknown. While new tools can enhance creators' workflows, the rapid influx of AI-generated content also threatens to dilute markets, suppress wages, and devalue human-made work. To date, few studies have assessed how AI will affect jobs and incomes for writers, actors, musicians, designers, and other

[5] Jack Hardinges, et al., *We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models* (December 13, 2023) Harvard Data Science Review, https://hdsr.mitpress.mit.edu/pub/xau9dza3/release/2.

creators. Further research will be necessary to fully understand the transformative effect that AI is having on California's creators.

A global report published by the International Confederation of Societies of Authors and Composers (CISAC and PMP Strategy) in December 2024 projected the emergence of a massive market for AI-generated content. The study estimated that AI-generated audiovisual content, including film and TV, could be worth €48 billion ($55.73 billion) by 2028. However, the report warned that up to 21 percent of creators' revenue in film and TV could be cannibalized as a result:

> In an unchanged regulatory framework, creators will actually suffer losses on two fronts: the loss of revenues due to the unauthorised use of their works by Gen AI models without remuneration; and replacement of their traditional revenue streams due to the substitution effect of AI-generated outputs, competing against humanmade works.[6]

The Concept Art Association describes the impact on their industry:

> Thousands of illustrators have found the entirety of their works, both personal and professional, used to train GenAI datasets without anyone having acquired consent, or having offered them any kind of credit or compensation. Vast quantities of data are routinely and indiscriminately scraped from across the internet to train GenAI models. These works serve as the "data fuel" for these models, and yet there is no requirement for GenAI developers to provide the lawful owners of these materials with any way to know if their specific works are contained within a dataset or not.

Meanwhile, the market for AI-generated music is projected to surge in the coming years. The CISAC study projected that the market for AI-generated music would grow to €64 billion ($74.3 billion) in 2028, with revenues for AI music platforms alone rising to about €4 billion ($4.64 billion). GenAI is forecast to claim roughly 20 percent of streaming music revenues by 2028, and AI-generated music has begun to regularly appear on the Billboard and Spotify charts.[7] One Guardian report found an entirely AI-created band – the Velvet Sundown – had amassed over one million Spotify streams.[8]

---

[6] *Study on the economic impact of Generative AI in the Music and Audiovisual industries* (November 2024) PMP Strategy & CISAC, https://www.cisac.org/services/reports-and-research/cisacpmp-strategy-ai-study.
[7] Aisha Down, *AI slop tops Billboard and Spotify charts as synthetic music spreads* (November 13, 2025) The Guardian, https://www.theguardian.com/technology/2025/nov/13/ai-music-spotify-billboard-charts.
[8] Lanre Bakare, *An AI-generated band got 1m plays on Spotify. Now music insiders say listeners should be warned* (July 14, 2025) The Guardian, https://www.theguardian.com/technology/2025/jul/14/an-ai-generated-bandgot-1m-plays-on-spotify-now-music-insiders-say-listeners-should-be-warned.

3. Creating a mechanism for identifying protected works

The author states the problem and pitches the solution:

> Generative artificial intelligence (GenAI) developers frequently use copyrighted materials to train new systems without crediting or compensating the owners of those materials. In order for copyright owners to exercise their rights over copyrighted materials, they must first know how their materials have been used. AB 412 increases transparency around the use of copyrighted materials to train GenAI by requiring developers to, upon receiving a request from a copyright owner, provide the owner with a list of copyrighted materials held by the owner that were used to train the system or model. Authors have a right to control and profit off of their own intellectual property, and AB 412 ensures they are able to do so.

The bill places a number of obligations on developers of GenAI models with respect to "covered materials" used to train them. That term is defined as materials registered, preregistered, or indexed with the United States Copyright Office pursuant to the federal Copyright Act of 1976 and sound recordings fixed before February 15, 1972, that are enforceable under federal law. Developers must first document any covered materials they *know* were used in training and make reasonable efforts to identify and document any other such training material. Then they are required to document the owner of the covered material.

To allow for proper identification, developers must make available information on their website that enables the generation of a fingerprint according to industry standards that is compatible with the covered materials used to train the relevant model. The website must include a mechanism that allows rights owners to submit a request for information about the developer's use of covered materials that allows the owner to provide specified information to the developer, including information sufficient to establish their identity and registration, preregistration, or index numbers and fingerprints for one or more covered materials. Upon receiving such a request, the developer must assess whether the fingerprint is likely to be present in the developer's dataset and provide the owner with the results.

The bill provides for a number of parameters intended to ensure there is not an undue burden on developers. This includes provisions that require response only when there is sufficient documentation and limit requests from rights owners to once per calendar quarter. The bill also includes a series of exemptions, such as where the relevant model is training exclusively using certain publicly available datasets, as provided, or where the covered materials are used pursuant to a licensing agreement.

Violations of the bill are subject to civil actions seeking actual or statutory damages and other relief. Rights owners are entitled to bring such actions if they have followed the stated guidelines and are not provided the required information.
SAG-AFTRA, a co-sponsor of the measure, explains the need for the bill:

> GenAI developers frequently use copyrighted materials to train new systems without crediting or compensating the owners of those materials. Unfortunately, copyright owners have no way of knowing whether a given copyrighted material was used to train a GenAI system or model. Some of those copyright owners left in the dark are our members and others they collaborate with in the entertainment industry. Protecting their rights is critical to protecting their work and their art. The transparency provided for in AB 412 will be invaluable in that endeavor.

4. The legal landscape

As noted above, should this bill pass out of this Committee, it will next be heard by the Senate Judiciary Committee. Therefore, this analysis will provide just a brief overview of copyright law. The U.S. Constitution grants Congress authority to establish exclusive copyright protections through Article I, Section 8, Clause 8, which empowers lawmakers to "promote the Progress of Science and useful Arts." Congress so used that power in establishing the Copyright Act of 1976 (the Copyright Act), which establishes copyright protections in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. (17 U.S.C. §§ 101 et seq.) It provides that all legal or equitable rights equal to those within the scope of the Copyright Act are governed exclusively by the Act, and no person is entitled to any such right or equivalent right in any such work under the common law or statutes of any State. (17 U.S.C. § 301.)

However, copyright law is "replete with escape valves: the idea-expression distinction; the general rule that facts may not receive protection; the requirement of originality; the legal standard for actionable copying; the limited duration of copyright; and . . . the defense of fair use."[9] When protected expression is copied, particularly in commercial contexts without attribution or transformation, questions of infringement may arise. The Copyright Act's fair use provision (17 U.S.C. § 107) identifies four factors for determining whether a given use of a copyrighted work is a "fair use":

---

[9] *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith* (2023) 598 U.S. 508, 550 ("Warhol"). Copyright exists from the moment the work is created. For copyright holders to sue for infringement, however, they must register the work with the United States Copyright Office, which administers an online catalog of registered copyright works, dating back to 1978, that enables the public to search for registered works by title, name, keyword, registration number, document number, or keyword command. (United States Copyright Office, Public Catalog, https://cocatalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First.)

> [T]he fair use of a copyrighted work . . . for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include —
>
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[10]

Courts have been clear that these four statutory fair use factors may not "be treated in isolation, one from another" but are "to be explored, and the results weighed together, in light of the purposes of copyright."[11]

GenAI systems are trained on large datasets that may include copyrighted works. Whether training models on copyrighted works without permission constitutes copyright infringement remains a significant unresolved issue. The question involves multiple overlapping considerations about the nature of AI training, the scope of copyright protection, and the application of existing legal doctrines to new technology. This includes resolving whether training AI models on copyrighted works constitutes fair use; whether models trained on copyrighted material that produce new-form outputs constitute infringement; and how commercialization of such models affects the economic framework copyright law establishes.

AI developers often assert that training on publicly available content qualifies as transformative fair use. Some academics and AI developers argue that training is highly transformative because the purpose differs fundamentally from the original works. While a novel entertains or informs readers, using that novel to train an LLM serves the distinct purpose of teaching a model about language patterns and knowledge representation. They also argue that the training process itself does not make copyrighted works available to end users or substitute for purchasing or licensing the original works. However, others are quick to point out that LLMs sometimes reproduce portions of training data, particularly when prompted in specific ways, suggesting that protected expression has been copied and retained. Some compare AI training to how humans learn by reading copyrighted works and absorbing information and patterns without infringing, provided they do not reproduce protected expression. However, many point out that, unlike humans, models can process millions of works

[10] *Bartz v. PBC* (2025) 787 F. Supp. 3d 1007, 1019-1020 ("*Bartz*").
[11] *Warhol*, at 550-551.

instantaneously, retain patterns with perfect consistency, and generate outputs at scale, making the human learning analogy imperfect.

Other experts and those in the creative industries argue these GenAI systems are developed and deployed for commercial profit, which weighs against fair use under the first factor of the statutory test. They point to the fact that model training involves copying entire works, often millions of them, which represents substantial taking under the third fair use factor.

A coalition of industry groups in opposition, including the California Chamber of Commerce, argues that the bill would interfere with pending litigation and runs afoul of existing copyright law and is preempted thereby:

> There are some who would argue that performing a computational analysis with publicly available works is a fair use and that there is no basis or grounds to sanction this activity with disclosure requirements. Courts have been actively considering whether the use of publicly available material constitutes a fair use of copyrighted material and this bill would impact pending litigation. Indeed, there are dozens of ongoing cases regarding the alleged use of copyrighted material to train artificial intelligence models. The Legislature should not place its thumb on the scales of justice and choose the winners of pending litigation.
>
> We suggest allowing these cases to proceed before determining whether a change in the law is necessary, especially now that the first cases are being decided and have provided important guidance regarding the application of copyright law to AI training. For example, in *Bartz v. Anthropic* (N.D. Cal., June 23, 2025), the court granted summary judgment in favor of Anthropic on its fair-use defense regarding the use of books for training its language model, while reserving other issues for further proceedings. Likewise, in *Kadrey v. Meta Platforms* (N.D. Cal., June 25, 2025), the court granted summary judgment for Meta on fair-use grounds. Although neither decision resolved all questions concerning AI training and copyright, both cases underscore that federal courts are actively developing the governing legal framework and that significant issues remain subject to further litigation and likely appellate review.
> …
> AB 412 is also expressly preempted under the U.S. Copyright Act. 17 U.S.C. Sec. 301, which provides that the subject matter and activities regulated by federal copyright law "are governed exclusively by this title" and that "no person is entitled to any such right or equivalent right in any such work under the…statutes of any State." While there are limited carveouts for state laws, none of those carveouts apply here.

These issues are currently working their way through the courts with many major cases left unresolved. A few key cases are described here.

In *Thomson Reuters Enter. Ctr. GmbH v. ROSS Intelligence Inc.* (D. Del. 2025) 529 F. Supp. 3d 303, 310, Thompson Reuters sued the defendant for copyright infringement, specifically for its use of Westlaw's editorial headnotes and its numbering system to train a competing AI tool, a narrow-purpose "predictive" system, rather than GenAI. The court found that the defendant's use constituted direct copyright infringement. The court rejected the fair use defense, citing commercial purpose, insufficient transformation, and potential market harm.

N.Y. Times Co. v. Microsoft Corp. (S.D.N.Y. 2023) 777 F. Supp. 3d 283, 300, involves allegations that articles and other protected works were used without authorization to train Microsoft and OpenAI's models, including ChatGPT. The plaintiffs' complaint claims copyright infringement related to both use at the training stage, where defendants use plaintiffs' works to train their LLMs, and the output stage, where defendants' LLMs generate outputs in response to user prompts that, according to the complaints, "regurgitate" plaintiffs' works. They argue that these uses potentially affect licensing revenue and market position. This case remains pending and is viewed as significant for publishing and news industries.

In *Bartz v. Anthropic PBC*, the uses at issue were Anthropic's downloading for free millions of copyrighted books in digital form from pirate sites on the internet and its purchase of copyrighted books in order "to amass a central library of 'all the books in the world' to retain 'forever.'" Anthropic then used a subset of these books to train various LLMs under development. The authors sued Anthropic for copyright infringement. On summary judgment, the issue was the extent to which any of the uses of the protected works qualified as a "fair use" under Section 107 of the Copyright Act.[12] The court's ruling on the motion found the use of the books at issue to train Anthropic's GenAI model, Claude, was "exceedingly transformative and was a fair use under Section 107 of the Copyright Act." However, it found Anthropic had "no entitlement to use pirated copies for its central library. Creating a permanent, general-purpose library was not itself a fair use excusing Anthropic's piracy." The parties are currently finalizing a $1.5 billion settlement as to the use of nearly 500,000 pirated works.[13]

---

[12] *Bartz*, 787 F. Supp. 3d at 1014.

[13] *Bartz v. Anthropic PBC* (2025) 2025 U.S. Dist. LEXIS 205531, *7;Blake Brittain, *US judge considers Anthropic's $1.5 billion settlement of authors' lawsuit* (May 14, 2026) Reuters, https://www.reuters.com/legal/government/us-judge-considers-anthropics-15-billion-settlement-authors-lawsuit-2026-05-14/.

5. Feasibility of the mechanisms and procedures called for by this bill

The industry coalition in opposition argues the bill's obligations are technically infeasible:

> Currently, there is no reliable, machine-readable database of copyright holders, making it impractical for even the smallest models to match or cross-reference copyright holders with data that was used to train a model, rendering this bill vastly impractical from a compliance standpoint alone. Consider, for example, the fact that a copyright owner can register and not use the copyright symbol, or that a person might use the copyright symbol and not register. Or that there can be multiple copyrighted elements on a single webpage (text, images, sound, video, ads with content of all those varieties).
>
> These issues are further complicated for works created prior to 1978, where registration information may be scattered across historical records and may be less uniformly digitized, requiring consultation across multiple archival sources. Determining ownership is therefore not simply a matter of conducting a search of publicly available records online. While the Copyright Office has expanded access to historical registration records, registration is not required for copyright protection, registration records do not necessarily identify the current copyright owner, and many works available on the internet contain little or no reliable ownership information. As a result, even where records exist, determining ownership remains a difficult and resource-intensive undertaking.
>
> We appreciate various attempts to address concerns around unregistered works with recent amendments, including attempts to address concerns that companies would have to somehow match every bit of copyrighted content to registered owner(s) via fingerprinting/hashing, and a mechanism that "allow[s] a rights owner to provide" certain documentation sufficient to establish the rights owner's identity; the physical or electronic signature of the rights owner or a third party authorized to act on behalf of the rights owner; and registration, preregistration, or index numbers and fingerprints for one or more covered materials. Despite those best efforts, however, these issues persist, and we fail to see any way in which the bill could adequately address these core concerns.
>
> Consider the introduction of concepts such as fingerprinting or hashing: part of the problem stems from a misperception that all material in a training data set will be used for training and that the data that is used for training will be in a format that will match a rights holder's "fingerprint"

> of specific copyrighted file. In reality, developers go through multiple steps to process, clean, and prepare data prior to training. This includes ensuring data is in the same format, dividing training data into different buckets for training, validation and testing, and the "tokenization" process. It is thus a highly complex question of how fingerprinting could be performed in a way that both matches rights holders to content and accurately reflects whether the content was used for AI model training.

The author points to the recent informational hearing held jointly by the Senate Judiciary and Assembly Privacy and Consumer Protection Committees in which testimony was taken that detailed how determining whether certain materials are within a large training dataset is, in fact, possible.

Writing in support, the Transparency Coalition.AI (TC) asserts:

> Given our technical subject matter expertise, TC also wishes to address two specific issues that have been raised by opponents to the bill: a) Opponents have questioned the technical feasibility of using fingerprints to match and identify copyrighted content that is part of the training data of an AI model. Fingerprinting is a technology that has been used for 30 years or more to match image, text, audio and video data in other application areas such as search and data management. As such, there should be no technical impediment to using the technology as planned for AB 412; b) Opponents have maintained that it is infeasible for them to document all "covered materials" in their possession because they "don't know what data was used." This is a false narrative since this documentation would be a simple side effect of the most rudimentary data engineering practices (that have been in place for close to 30 years) that are considered as industry norms and best practices today. Maintaining this documentation would also facilitate compliance with CA AB 2013 (as required by law from Jan 1, 2026).

6. Additional stakeholder positions

The Animation Guild writes in support:

> The Animation Guild is a labor union that represents more than 11,000 artists, technicians, writers, and production workers in the animation industry, most of whom live and work in California. The California-based animation industry alone drives billions of dollars in ticket sales, merchandise, and tourism. The artificial intelligence industry relies on scraped data produced by the hard work of our members, often to create products that directly seek to professionally displace them. Without the transparency required by AB-412, our members will be at a serious

> disadvantage to understanding how and where their work is being used and exercising their rights.

The Professional Photographers of America write in support:

> By fostering accountability, AB 412 protects the intellectual property of photographers and other creators, ensuring their contributions are respected in the age of AI innovation. This legislation does not stifle innovation or technological progress, rather it will ensure longstanding copyright protections are maintained to protect thousands of small businesses within the great state of California.

The Consumer Technology Association (CTA) writes in opposition:

> CTA raised concerns about AB 412 in 2025, and those concerns remain. Since then, technology has advanced quickly. Content creators now have more tools than ever to control where and how their works are used. Before advancing a one-size-fits-all mandate, the committee should consider whether technology solutions can better address these concerns than legislation. Requiring companies to identify and catalog all potential copyright holders in AI training datasets is unrealistic at scale, especially for startups and small businesses. This burden compounds further when applied to internal tools and incremental model modifications — uses that were never the intended target of copyright transparency legislation.

7. Amendments

In response to the above concerns, the author has agreed to a number of amendments that significantly narrow the bill and remove prescriptive technical language. The bill as agreed to be amended simply requires a developer to make available on its website a mechanism allowing a rights owner to request information about the developer's use of the rights owner's covered materials. The mechanism must allow a rights owner to provide the developer with specified information. The developer in turn must, within 30 days, assess whether the developer used the relevant covered materials to develop the GenAI model and notify the rights owner accordingly.

SUPPORT

Concept Art Association (co-sponsor)
SAG-AFTRA (co-sponsor)
American Federation of Musicians, Local 7
American Society for Collective Rights Licensing
Animation Guild, IATSE Local 839
Arte Es Ética - Artivismo Y Autorías De Habla Hispana

Asifa-Hollywood
California Arts Advocates
California Federation of Labor Unions, AFL-CIO
Charlotte Ann Voice-overs, LLC
Entertainment Union Coalition
Ethovox INC.
European Guild for Artificial Intelligence Regulation
Facilitating Performance
Fantagraphics
Graphic Artists Guild
Heart of LA Democratic Club
Heavy Metal
IATSE Local 800
Joanne's English Club
Lava Latin Association of Voice Actors
Los Angeles County Democratic Party
Lucita INC
National Association of Voice Actors
Nippon Anime and Film Culture Association
NorCal Voices
Professional Photographers of America
Santa Monica Democratic Club
SJSU Animation & Illustration Program
Society of Illustrators of Los Angeles
Soundbox:studio Group
Transparency Coalition.AI
17,131 individuals

OPPOSITION

AI Salon
American Innovators Network
Bay Area Council
Bizfed Central Valley
Business Software Alliance
Calbroadband
California Chamber of Commerce
California Credit Union League
Chamber of Progress
Civil Justice Association of California (CJAC)
Computer & Communications Industry Association
Consumer Technology Association
Developers Alliance
Ecomback

Electronic Frontier Foundation
Engine Advocacy
Insights Association
Internet Coalition
Los Angeles County Business Federation (BIZFED)
Re:create
San Diego Regional Chamber of Commerce
San Jose Chamber of Commerce
San Mateo County Economic Development Association (SAMCEDA)
Silicon Valley Leadership Group
Software Information Industry Association
Techca
Technet
Valley Industry & Commerce Association

RELATED LEGISLATION

AB 2713 (Wicks, 2026) updates the California AI Transparency Act to allow a large online platform to permit users to download content provenance data directly from the platform, and to prohibit a large online platform from knowingly stripping system provenance data or a digital signature from user-downloaded content. AB 2713 is currently pending referral in the Senate.

AB 2013 (Irwin, Ch. 817, Stats. 2024) requires developers of GenAI systems or services that are made available for Californians to use to post on their website documentation regarding the data used to train the system or service, including high-level summaries of the datasets used.

PRIOR VOTES:

Assembly Floor (Ayes 45, Noes 16)
Assembly Judiciary Committee (Ayes 9, Noes 1)
Assembly Privacy and Consumer Protection Committee (Ayes 10, Noes 3)

****