SENATE JUDICIARY COMMITTEE
Senator Thomas Umberg, Chair
2025-2026 Regular Session

AB 412 (Bauer-Kahan)
Version: June 11, 2026
Hearing Date: June 23, 2026
Fiscal: No
Urgency: No
AWM

SUBJECT

Generative artificial intelligence: training data: copyrighted materials

DIGEST

This bill requires Generative AI (GenAI) developers to document covered material, as defined, used to train their models and to facilitate a process through which a copyright holder can determine whether their covered material was used in a model's training.

EXECUTIVE SUMMARY

The use of copyrighted works to train AI models has sparked significant legal and ethical debate. Given that GenAI can produce text, images, video, or audio that emulates the expressive works used to train it, potentially destroying the market for the artists whose work was used to train the GenAI, artists and rightsholders are sounding the alarm. The question of how far the Copyright Act extends in the GenAI training context, however, is unresolved. Key questions include whether training on copyrighted material without permission constitutes fair use or copyright infringement, particularly when models can sometimes reproduce verbatim portions of their training data. There are also questions about transparency: many users and creators are unaware of which copyrighted materials were included in training datasets.

This bill is intended as a transparency measure for copyright holders. Under the bill, the developer of a GenAI model commercially available in California must provide a mechanism through which a rightsholder can seek information about whether their materials were used to train the model. The developer must conduct an assessment and respond to the request within 30 days, provided that the rightsholder's request included the requisite information to identify their copyrighted works. If a developer fails to comply, the rightsholder can bring a civil action for actual damages or a civil penalty of up to $1,000 per violation. The bill exempts from its requirements GenAI models that meet specified criteria relating to their training or purpose, including if the training datasets are already publicly available or do not include copyrighted material, or if the

GenAI model is developed and used by a university or government entity and used solely for noncommercial academic or government research. The author has agreed to amend the bill to give an operator a 30-day right to cure before a rights owner may bring a civil action for a violation of the bill.

This bill is sponsored by the Concept Art Association, the National Association of Voice Actors, and the Screen Actors Guild-American Federation of Television and Radio Artists (SAG-AFTRA), and is supported by a number of creators' associations, advocacy organizations, labor groups, and over 23,000 individuals. This bill is opposed by opposed by a number of industry associations, including Technet and CalBroadband. The Senate Privacy, Digital Technologies, and Consumer Protection Committee passed this bill with a vote of 6-2.

PROPOSED CHANGES TO THE LAW

Existing constitutional law:

1) Provides that the United States Congress shall have 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) Provides that the U.S. Constitution, and the Laws of the United States, are the supreme law of the land. (U.S. Const., art. VI, cl. 2.)

Existing federal law:

1) Establishes the Copyright Act of 1976 (the Act), which sets forth the requirements for an author of a creative work to obtain a copyright in that work. (17 U.S.C. §§ 101 et seq.)

2) Provides that the owner of a copyright under the Act has the exclusive rights to do and to authorize any of the following:
    a) To reproduce the copyrighted work in copies or phonorecords.
    b) To prepare derivative works based upon the copyrighted work.
    c) To distribute copies or phonorecords of the copyrighted work to the public by sale or transfer of ownership, or by rental, lease, or lending.
    d) In the case of literary, musical, dramatic, and choreographic works, pantomimes, motion pictures and other audiovisual works, to perform the copyrighted work publicly.
    e) In the case of literary, musical, dramatic, and choreographic works, pantomimes, and pictorial, graphic, or sculptural works, including the images of a motion picture or other audiovisual work, to display the copyrighted work publicly.

f) In the case of sound recordings, to perform the copyrighted work publicly by means of a digital audio transmission. (17 U.S.C. § 106.)

3) Provides, notwithstanding 2), that fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright; and that in determining whether the use made of a work in any particular case is a fair use, the factors to be considered include:
   a) The purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes.
   b) The nature of the copyrighted work.
   c) The amount and substantiality of the portion used in relation to the copyrighted work as a whole.
   d) The effect of the use upon the potential market for, or value of, the copyrighted work. (17 U.S.C. § 107.)

4) Provides that, beginning January 1, 1978, all legal or equitable rights that are equivalent to any of the exclusive rights within the general scope of copyright as specified in 2) in works of authorship that are fixed in a tangible medium of expression and come within the scope of copyright, as defined, are governed exclusively by the Act; as of that date, 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(a).)

5) Provides, notwithstanding 4), that nothing in the Act annuls or limits any rights or remedies under the common law or statutes of any state with respect to the following:
   a) Subject matter that does not come within the subject matter of copyright, including works of authorship not fixed in any tangible medium of expression.
   b) Any cause of action arising from undertakings commenced before January 1, 1978.
   c) Activities violating legal or equitable rights that are not equivalent to any of the exclusive rights within the general scope of copyright as specified by 2).
   d) State and local landmarks, historic preservation, zoning, or building codes relating to protected architectural works, as specified. (17 U.S.C. § 301(b).)

6) Provides that any person who violates any of the exclusive rights of a copyright owner, as defined, is an infringer of the copyright, and that the legal or beneficial owner of the copyright may bring a civil action for infringement and obtain damages and injunctive relief. (17 U.S.C. §§ 501-504.)

Existing state law establishes the California AI Transparency Act, which 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) Defines the following terms:
    a) "Artificial intelligence" or "AI" means an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.
    b) "Covered material" means either (1) a material registered, preregistered, or indexed with the United States Copyright Office under the Act, or (2) a sound recording fixed before February 15, 1972, enforceable under specified federal law. "Covered material" does not mean a material that is in the public domain.
    c) "Rights owner" means the owner of a covered material.
    d) "Developer" means a business, person, partnership, corporation, or other entity that designs, codes, produces, or substantially modifies a GenAI model and that either (1) uses the GenAI model commercially in California, or (2) makes the GenAI model available to Californians for reasonably foreseeable commercial use.
    e) "Generative artificial intelligence" or "GenAI" means an artificial intelligence system that can generate derived synthetic content, including text, images, video, and audio, that emulates the structure and characteristics of the system's training data.

2) 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.

3) Requires the mechanism in 2) to allow a rights owner to provide all of the following to the developer:
    a) Documentation sufficient to establish the rights owner's identity.
    b) The physical or electronic signature of the rights owner or a third party authorized to act on their behalf.
    c) Registration, preregistration, or index numbers for one or more of the rights owner's covered materials.
    d) Any additional information specified by the developer that is reasonably necessary to comply with this bill, except that a rights owner shall not be required to transmit a copy of a covered material in a form suitable for

training, fine-tuning, or otherwise developing a GenAI model to a developer in order to receive information about the developer's use of a covered material.

4) Requires a developer to document and retain any requests received from rights owners pursuant to 2) for as long as the developer uses the GenAI model commercially in California or makes the GenAI model available to Californians for reasonably foreseeable commercial use, whichever is longer, plus five years.

5) Requires a developer, within 30 days of receiving a request for information from a rights owner under 2), to do both of the following:
    a) Assess whether the developer used the rights owner's covered materials to develop the GenAI model.
    b) Provide the rights owner with a list of any identified covered materials.

6) Requires the assessment in 5)(a) to be all of the following:
    a) Designed to identify or extract substantially similar copies of the covered materials in training datasets or other records maintained by the developer, including through the use of approximate content fingerprints or functionally equivalent technical measures where appropriate.
    b) Robust to minor variations in covered materials, including changes in file format, resolution, cropping, resizing, excerpting, or other modifications that do not change the expressive meaning or functional content of the materials.
    c) Appropriate to the format of the covered material, including the text, images, audio, video, or other protected works.
    d) Applied to all training datasets and other records maintained by the developer that are reasonably likely to contain information related to the request.

7) Provides that a developer's collection, use, retention, and sharing of information from a rights owner pursuant to 2) 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.

8) Provides that each day after the 30-day period in 5) that a developer fails to provide a rights owner with the information required under this bill constitutes a discrete violation.

9) Provides that a developer is not required to respond to a request that is either of the following:
    a) Not accompanied by documentation sufficient to establish the rights owner's identity.

b) Made in violation of 10).

10) Provides that a rights owner, or any person acting on their behalf, shall not submit 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; a single request, however, may pertain to multiple covered materials.

11) Provides that a rights owner that has complied in good faith with 10), and that is not provided with the information as required by this bill, may bring a civil action against the developer for any of the following:
    a) $1,000 per violation or actual damages, whichever is greater.
    b) Injunctive or declaratory relief.
    c) Reasonable attorney's costs and fees.
    d) Any other relief the court deems appropriate.

12) Provides that this bill shall not apply to a GenAI model that is any of the following:
    a) Trained exclusively using datasets the developer makes publicly available at no cost to users of the developer's website.
    b) Trained exclusively using datasets a third party makes publicly available at no cost to users, as disclosed by the developer pursuant to the California AI Transparency Act.
    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.

13) Provides that nothing in this bill shall be construed to limit liability on the provider of a telecommunications service, information service, or cable service, as those terms are defined in Section 153 of Title 47 of the United States Code, for content provided by another person, to the extent the provider is not acting as the developer of a GenAI model.

COMMENTS

1. Author's comment

According to the author:

> 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.

2. Background on GenAI and GenAI training process

As explained by the Senate Privacy, Digital Technologies, and Consumer Protection Committee's analysis of this bill:

> 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]

[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 links in this analysis are current as of June 18, 2026.

> 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. Background on the Copyright Act

The United States Constitution grants Congress "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."[3] The "useful arts" are currently secured by the Copyright Act of 1976 (the Act).[4] Section 106 of the Act sets forth the exclusive rights enjoyed by the author of the work, or their assignee, including the right to reproduce, and prepare derivative works based on, the copyrighted work.[5]

The Act also provides, however, that a copyright holder may not prevent "fair use" of their copyrighted work.[6] The fair use doctrine is an "equitable rule of reason that permits courts to avoid rigid application of the copyright statute when, on occasion, it would stifle the very creativity which that law is designed to foster."[7] Fair uses contemplated by the Act are "purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research."[8] To determine whether a specific act counts as fair use, a court will apply a four-part test weighing "the purpose and character of the use; the nature of the work; the amount and substantiality of the portion of the work used; and the effect of the use on the potential market for, or value of, the copyrighted work."[9]

4. Pending litigation relating to GenAI training and copyright

As discussed above in Comment 2, GenAI requires heretofore unimagined amounts of data to create even a passably useful model. Most, if not all, of the largest commercially available LLMs were trained on copyrighted material without the rightsholders'

---

[2] Schreiner, *GPT-4 architecture, datasets, costs and more leaked* (Jul. 11, 2023) The Decoder, https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.
[3] U.S. Const., art. I, § 8, cl. 8.
[4] 17 U.S.C. §§ 101 et seq.
[5] *Id.*, § 106.
[6] *Id.*, § 107.
[7] *Google LLC v. Oracle America, Inc.* (2021) 593 U.S. 1, 18 (internal quotation marks omitted).
[8] 17 U.S.C. § 107.
[9] *Google LLC*, *supra*, 593 U.S. at p. 19.

permission.[10] Whether, and to what extent, the Act permits training a GenAI model with copyrighted works without permission from the rightsholder is the source of dozens of pending lawsuits. Committee staff are aware of only three cases in which a trial court ruled directly on this issue. These three cases are summarized below.

a. Bartz v. Anthropic PBC

Bartz was filed by a group of book authors whose copyrighted works were copied by Anthropic in connection with the training of its Claude LLM.[11] The plaintiffs sought to represent a class of book authors whose works were used to train Claude.[12]

"From the start, Anthropic had many places from which it could have purchased books, but it preferred to steal them to avoid 'legal/practice/business slog,' as cofounder and chief executive officer Dario Amodei put it."[13] Between 2021 and 2022, Anthropic downloaded over seven million copies of books from known pirating sites.[14] In 2024, Anthropic switched to purchasing millions of print books—often used—stripping them down, cutting their pages down to size, and scanning the books into digital form.[15] Anthropic did not use every single digital version of a book it acquired or copied to train Claude; some copies it merely retained for potential future use.[16]

Anthropic filed a motion for summary judgment, arguing that all of its uses of plaintiffs' copyright material were fair use and therefore not prohibited by the Act.[17] The court agreed with Anthropic that "the purpose and character of using copyrighted works to train LLMs to generate new text was quintessentially transformative…Anthropic's LLMs trained upon works not to race ahead and replicate or supplant them—but to turn a hard corner and create something different."[18] The court emphasized that this ruling related to inputs only; to the extent that Claude produced outputs that infringed on the authors' works, the authors could bring such a case.[19]

That was not, however, the end of the court's inquiry. Anthropic also made and retained copies of all of the books it acquired, whether or not they ended up in the Claude training sets.[20] Copying a copyrighted work is one of the rights reserved to the

---

[10] *E.g.*, Edwards, *OpenAI says it's "impossible" to create useful AI models without copyrighted material* (Jan. 9, 2024) ars technica, https://arstechnica.com/information-technology/2024/01/openai-says-its-impossible-to-create-useful-ai-models-without-copyrighted-material/.
[11] *Bartz v. Anthropic PBC* (N.D. Cal. 2025) 787 F.Supp.3d 1007, 1014.
[12] *Id.* at p. 1018.
[13] *Id.* at p. 1015.
[14] *Ibid.*
[15] *Id.* at p. 1016.
[16] *Id.* at p. 1018.
[17] *Ibid.*
[18] *Id.* at p. 1021-1022.
[19] *Id.* at p. 1021.
[20] *Id.* at p. 1022.

rightsholder under the Act, and "Anthropic seem[ed] to believe that because some of the works it copied were sometimes used in training LLMs, Anthropic was entitled to take for free all the works in the world and keep them forever with no accounting. There is no carveout, however, from the Copyright Act for AI companies."[21] The court ruled that there was no infringing use for the physical books Anthropic purchased, digitized, and destroyed, because "that format change added no new copies."[22] But for the pirated works, the court rejected Anthropic's argument that, because the intended use for the copies was a fair use, Anthropic did not violate the Act by making unauthorized copies.[23] "Such piracy of otherwise available copies is inherently, irredeemably infringing even if the pirated copies are immediately used for the transformative use and immediately discarded."[24] Anthropic was, thus "wrong to suppose that so long as you create an exciting end product, every 'back-end step, invisible to the public' is excused."[25] The court thus denied Anthropic's summary judgment motion with respect to the downloaded pirated copies.[26]

Rather than go to trial on whether it infringed the copyrights of the authors whose works it pirated, Anthropic settled the case and established a $1.5 billion fund to pay authors whose works it infringed.[27] Each author whose work was in the class of works pirated by Anthropic was expected to get around $3,000; after large numbers of authors opted out of the class, however, the court asked for further briefing on the proposed settlement terms, including whether plaintiffs' counsel's requested $320 million fee award was too high.[28]

b. Kadrey v. Meta

Similar to *Bartz*, *Kadrey* was filed by 13 book authors who sued Meta for alleged infringement for Meta's use of copies of their books downloaded from online "shadow libraries" to train Meta's Llama LLM.[29] Meta tried to negotiate licensing deals with several book publishers to use published works as training data, but "as negotiations proceeded, Meta realized that licensing would be more difficult than anticipated."[30] So

---

[21] *Ibid.*
[22] *Id.* at p. 1023.
[23] *Id.* at p. 1025.
[24] *Ibid.*
[25] *Id.* at p. 1027; *see also id.* at p. 1033 ("Every factor points against fair use. Anthropic employees said copies of works (pirated ones too) would be retained 'forever' for 'general purpose' even after Anthropic determined they would never be used for training LLMs. A separate justification was required for each use. None is even offered here except Anthropic's pocketbook and convenience.").
[26] *Id.* at p. 1033.
[27] *See Bartz, et al. v. Anthropic PBC* Copyright Settlement Page, Frequently Asked Questions, https://www.anthropiccopyrightsettlement.com/faq.
[28] Belanger, *Anthropic's $1.5B copyright settlement is getting messy as judge delays approval* (May 15, 2026) ars technica, https://arstechnica.com/tech-policy/2026/05/authors-fight-for-higher-payouts-from-anthropics-1-5b-copyright-settlement/.
[29] *Kadrey v. Meta* (N.D. Cal. 2025) 788 F.Supp.3d 1026, 1036.
[30] *Ibid.*

instead, with the blessing of Meta CEO Mark Zuckerberg, Meta turned to the shadow libraries.[31]

Meta did not dispute that the plaintiffs had a prima facie case for infringement on their right to reproduce their works under the Act, but argued, and sought summary judgment on the basis that, its reproduction constituted fair use.[32] The court agreed and reached the same conclusion as *Bartz*: using copyrighted materials to train an LLM, at least on the basis of the facts of the case and the arguments made by plaintiffs, was fair use.[33] Judge Chhabria, however, seemed less than thrilled with the result, stating at the end of the order:

> In cases involving uses like Meta's, it seems like the plaintiffs will often win, at least where those cases have better-developed records on the market effects of the defendant's use. No matter how transformative LLM training may be, it's hard to imagine that it can be fair use to use copyrighted books to develop a tool to make billions or trillions of dollars while enabling the creation of a potentially endless stream of competing works that could significantly harm the market for those books. And some cases might present even stronger arguments against fair use. For instance, as discussed above, it seems that markets for certain types of works (like news articles) might be even more vulnerable to indirect competition from AI outputs. On the other hand, though, tweak some facts and defendants might win. For example, using copyrighted books to train an LLM for nonprofit purposes, like national security or medical research, might be fair use even in the face of some amount of market dilution…[34]

He suggested that "perhaps the plaintiffs could even have made a strong enough showing to win on the fair use issue at summary judgment. But the plaintiffs presented no meaningful evidence on market dilution at all. Absent such evidence and in light of Meta's evidence, the fourth factor can only favor Meta."[35]

c. Thomson Reuters v. Ross Intelligence Inc.

Thomson Reuters owns the legal research platform Westlaw.[36] Westlaw hosts exact copies of judicial orders and opinions, statutes, law journals, and treatises, to which it adds its own editorial content in the form of its Key Number System, which creates a taxonomy of legal issues with which researchers can sort cases and other materials.[37]

---

[31] *Id.* at p. 1041.
[32] *Id.* at p. 1042.
[33] *Id.* at p. 1060.
[34] *Id.* at pp. 1059-1060.
[35] *Id.* at p. 1060.
[36] *Thomson Reuters Enterprise Centre GMBH v. Ross Intelligence Inc.* (D. Del. 2025) 765 F.Supp.3d 382, 390.
[37] *Ibid.*

Ross is a legal-AI startup that intended to compete with Westlaw.[38] Ross "asked to license Westlaw's content" to train its AI, but Westlaw refused.[39] Ross then purchased in bulk, from a third party, legal memos built off of Westlaw headnotes. "When Thomson Reuters found out, it sued Ross for copyright infringement."[40]

In 2023, the court largely denied Thomson Reuters's motion for summary judgment to foreclose Ross's fair use argument.[41] Two years later, however, the court reconsidered its ruling and granted summary judgment for Thomson Reuters on the question of Ross's fair use.[42] One key difference between the judge's reasoning in *Thomson Reuters* as opposed to *Bartz* and *Kadrey* is that the *Thomson Reuters* judge did not believe Ross's use of Westlaw headnotes was transformative.[43] Ross argued that its use of the headnotes was akin to existing law holding that copying computer code, for the purpose of developing more innovative code, was fair use, but the court rejected this argument.[44] Instead, the court concluded that Ross used the headnotes simply "to make it easier to develop a competing research tool."[45] The court also found that the fourth fair use factor—the effect of the unauthorized use on the market for the original—weighed in favor of Thomson Reuters, because Ross's goal was to create a commercial competitor to Westlaw.[46]

Ross asked the court to certify the ruling for an interlocutory appeal to the United States Court of Appeal for the Third Circuit, and the court granted the request.[47] The Third Circuit heard oral argument in this case on June 11, 2026.

5. <u>This bill requires a GenAI developer to provide a mechanism through which a copyright holder can request information about whether their work was used to train the GenAI model</u>

This bill gives rightsholders a mechanism to request information from the developer of a GenAI model commercially available in California regarding the developer's use of the rightsholder's copyrighted materials. A covered developer must make the mechanism available on its website and must permit the rightsholder to provide the developer with information sufficient to identify themselves as the rightsholder and the registration, preregistration, or index numbers for one or more of the rightsholder's materials. A rightsholder can submit only one request per calendar quarter concerning

---

[38] *Ibid.*
[39] *Ibid.*
[40] *Id.* at p. 391.
[41] *Ibid.*
[42] *Id.* at p. 397.
[43] *Id.* at p. 398.
[44] *Id.* at p. 399.
[45] *Ibid.*
[46] *Id.* at p. 400.
[47] *Thomson Reuters Enterprise Centre GMBH v. Ross Intelligence Inc.* (D. Del., May 23, 2025) Case No. 1:20-cv-613-SB, 2025 WL 1488015.

the same GenAI model, but a single request may address more than one copyrighted work.

When a rightsholder submits a complete and compliant request, the developer must determine whether the rightsholder's copyrighted materials were used to train the developer's GenAI model, including determining whether exact or substantially similar copies of the materials were used, and whether or not they were in different file formats, resolutions, or other modifications that don't affect the expressive meaning or functional content of the materials. The assessment must be made of all training datasets and other records maintained by the developer that are reasonably likely to contain information related to the rightsholder's request. The developer must then provide the rightsholder with a list of the rightsholder's copyrighted materials located in the assessment. The assessment and response must be completed within 30 days of receipt of the request.

The bill establishes a cause of action for a rightsholder who submitted a request and did not receive a response as required. The rightsholder may bring a civil action against the developer for $1,000 per violation or actual damages, whichever is greater; injunctive or declaratory relief; reasonable attorney's fees and costs; and any other relief the court deems appropriate. The bill specifies that each day that a response is delayed, after the required 30-day response window, constitutes a separate violation of the bill. The author has agreed to amend the bill to provide a right to cure period for an operator whom a rightsholder believes has violated the bill's provisions; the amendments are set forth in Comment 7, below.

The bill exempts GenAI models trained on publicly available datasets, as specified, not trained on copyrighted materials, or trained on copyrighted materials which the developer either owned or licensed; and GenAI models developed and used solely by universities or government for academic or research purposes, or developed and used exclusively for the operation of aircraft in the national airspace.

This bill is intended as a transparency measure for copyright holders whose works may have been used to train GenAI models. While AB 2013 (Irwin, Ch. 817, Stats. 2024) requires GenAI developers to disclose "high-level summar[ies]" of their training datasets,[48] the only way the holder of a copyright can determine whether their specific work was used to train an AI model is to file a lawsuit. Given that the developers of the largest GenAI models are also some of the wealthiest corporations in the world, a lawsuit is simply out of the question for many artists.

Opponents of the bill argue that this bill will not provide artists with useful information because the use of a copyrighted work is fair use. As discussed in Comment 4, this is a rapidly evolving area of law. Moreover, as *Bartz* and *Kayden* indicate, the circumstances

[48] Civ. Code, § 3111.

surrounding the acquisition of materials, and the particulars of the model and the potential for affecting the market for works, are all fact-intensive inquiries that make a blanket "training is fair use" rule impossible.

The other issue is outputs, i.e., the materials generated by a GenAI trained on copyrighted materials. Lawsuits relating to infringing outputs are also ongoing, though some GenAI developers seem a little more cautious about their chances of success. For example, when OpenAI released its Sora 2 generative video model, OpenAI allowed copyright holders to opt out of having their characters appear in Sora-generated videos.[49]

Unlike patent rights, copyrights can be granted to substantially similar, or even identical, works if the works were created truly independently of each other.[50] Moreover, mere evidence of similarity between two works is insufficient for a copyright holder to demonstrate that a substantially similar work is infringing; the rightsholder must have either direct evidence that the alleged infringer copied the work, or show that the alleged infringer had direct access to the work and substantial evidence of copying.[51] "A showing of access requires 'either evidence of a chain of events between the plaintiff's work and defendants' access to that work or evidence that the plaintiff's work has been widely disseminated.' "[52] For an independent artist whose art was not widely published or known, faced with an AI-generated image that is similar, but not identical to, their works, the only way to determine whether it's a coincidence or potential infringement is to file a prohibitively expensive lawsuit.

Finally, opponents of the bill make a number of arguments about the technical feasibility and cost of this bill's requirements. The Senate Privacy, Digital Technologies, and Consumer Protection Committee considered this bill from a technical standpoint and amended the bill significantly in response to those concerns. That Committee also considered this bill from an overall tech standpoint. The remaining issue for this Committee is whether this bill, as opponents argue, is preempted by the Act. This issue is discussed in Comment 6, below.

6. <u>The preemption question</u>

a. Express preemption

The Act includes an express preemption clause stating that "all legal or equitable rights that are equivalent to any of the exclusive rights within the general scope of copyright

---

[49] *E.g.*, Barkin, *OpenAI's "Sora" Sparks Copyright and Fair Use Debate* (Oct. 24, 2025) Harvard Law School Journal of Sports & Entertainment Law, https://journals.law.harvard.edu/jsel/2025/10/openais-sora-sparks-copyright-debate/. The Act does not, in fact, include an "it's not infringement if you didn't tell me not to" exception. (*See generally* U.S.C. Title 17.)
[50] *E.g.*, *Biani v. Showtime Networks, Inc.* (9th Cir. 2025) 153 F4th 957, 963.
[51] *Ibid.*
[52] *Ibid.*

as specified by section 106 in works of authorship that are fixed in a tangible expression…are governed exclusively by [the Act]" for works created on or after January 1, 1978.[53] For those works, "no person is entitled to any such right or equivalent right in such work under the common law or statutes of any state."[54] The Act does not, however, preempt "activities violating legal or equitable rights that are not equivalent to any of the exclusive rights within the general scope of copyright as specified by section 106."[55] "In short, a state-law claim is preempted if (1) the work is within the scope of the subject matter of copyright as specified in 17 U.S.C. sections 102 and 103; and (2) the rights granted under state law are equivalent to any exclusive rights within the scope of federal copyright as set out in 17 U.S.C. section 106."[56]

This bill probably satisfies the first element for express preemption: by definition, the bill's creates a right for copyright holders with respect to their copyrighted works.

On the second element, however, this bill does not appear to give rightsholders any equivalent rights to those granted in Section 106. The leading Nimmer on Copyright Treatise states:

> [I]f qualitatively other elements are required, instead of, or in addition to, the acts of reproduction, performance, distribution, or display, in order to constitute a state-created cause of action, then the right does not lie "within the general scope of copyright," and there is no preemption.[57]

This bill does not create any Section 106-adjacent rights—the bill does not purport to prevent training on copyrighted works, require payment for training on copyrighted works, or otherwise alter a developer's use of the data—and has qualitatively additional elements in the form of the request submission, timeframes, and so on. It therefore does not appear that this bill is preempted under the Act's express preemption clause.

b. Implied preemption

In addition to the Act's express preemption clause, courts have found state law claims preempted "where state law stands as an obstacle to the accomplishment and execution of the full purposes and objectives of Congress."[58] "What constitutes a 'sufficient obstacle' is a 'matter of judgment, to be informed by examining the federal statute as a whole and identifying its purpose and intended effects.' "[59]

---

[53] *Id.*, § 301(a).
[54] *Ibid.*
[55] *Id.*, § 301(b).
[56] *R.W. Beck, Inc. v. E3 Consulting, LLC* (10th Cir. 2009) 577 F.3d 1133, 1146 (cleaned up).
[57] 1 Nimmer on Copyright § 1.14 (2026).
[58] *Ryan v. Editions Ltd. West, Inc.* (9th Cir. 2015) 786 F.3d 754, 761 (internal quotation marks omitted).
[59] *In re Jackson* (2d Cir. 2020) 972 F.3d 25, 34.

As discussed above, this bill does not directly provide copyright holders with a new avenue to control the use of, or vindicate the misuse of, their copyrights. Opponents of the bill, however, argue that, because a separate portion of the Act provides a pre-litigation subpoena power to rightsholders in certain situations, the creation of a state law transparency measure should be understood to conflict with the Act as a whole. The Ninth Circuit Court of Appeals has not clearly articulated a test for conflict preemption under the Act; it is therefore unknown whether this measure would be considered an "obstacle" to the Act's overall purpose.

7. Amendments

As noted above, the author has agreed to amend the bill to add a right-to-cure period for an operator. The amendments are set forth below, subject to any nonsubstantive changes the Office of Legislative Counsel may make.

Amendment

On page 7, modify lines 16-24 as follows:

3117. ***(a) Subject to subdivision (b), a*** rights owner that has complied in good faith with Section 3116.5 and that is not provided with the information as required by this title may bring a civil action against the developer for any of the following:

(1) One thousand dollars ($1,000) per violation or actual damages, whichever is greater.

(2) Injunctive or declaratory relief.

(3) Reasonable attorney's costs and fees.

(4) Any other relief the court deems appropriate.

(b) Before initiating an action under subdivision (a), a rights owner shall provide the developer 30 days' notice of intent to file suit. During this period, the developer may cure any of the developer's violations of this title pertaining to the rights owner.

8. Arguments in support

According to the Concept Art Association:

California has always been on the forefront of policies that can lead the way for America, and the AI Copyright Transparency Act is another great representation of this. AB-412 presents California with the opportunity to continue leading by

example, as it empowers artists and creators to exercise their rights as copyright owners. Federal law provides copyright holders with a variety of exclusive rights over their works, including the right to reproduce, distribute and display copyrighted materials.

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…

Frankly, GenAI developers should want to provide this information. First, this gives developers a way to assure rights holders if their materials are not contained within a model's dataset. Second, if by chance copyrighted materials are contained within a model's dataset, it presents both sides with an opportunity to remedy the situation and make a contract or deal for their use.

There is a great need for transparency in this realm on all sides. Artists and creators must have the ability to know, protect, and enforce their copyrights. Similarly, GenAI developers must have the ability to protect themselves from unwarranted claims that their datasets contain copyrighted works.

9. Arguments in opposition

According to a coalition of the bill's opponents:

To be clear, we take our members' copyright interests seriously, especially considering the diverse ways in which different members may be affected by various copyright laws. Our focus is on ensuring fairness and supporting compliance with the requirements placed on all parties, rather than favoring one side over the other within our membership. With that in mind, we have serious concerns about AB 412 because its requirements are not feasible as a practical matter, create significant risks to proprietary and sensitive information, impose substantial compliance burdens that disproportionately disadvantage smaller AI companies and startups, and undermine recently enacted legislation addressing AI transparency. Preemption concerns aside, we ultimately feel that this bill is neither necessary due to the recent passage of AB 2013 (Irwin, Chapter 817, Statutes of 2024), nor feasible as a practical matter, and are concerned about its overall impact on California businesses and economy due to statutory penalties, disclosures of proprietary or otherwise sensitive information, the creation of a state-compelled prelitigation discovery regime, and unnecessarily creates uncertainty while key legal issues remain under active judicial consideration, among other things. It also

becomes increasingly unnecessary as copyright owners increasingly have access to technical tools and industry mechanisms that can be used to signal preferences regarding the use of their publicly available content by automated systems…

It is unclear to us why **A**B 412 needs to add new statutory damages when copyright law already provides an extremely generous and advantageous statutory damages regime for plaintiffs. Despite the inclusion of a good faith compliance requirement, this becomes particularly problematic here where thousands of potential plaintiffs can claim to be the owner of a copyrighted work the moment they make a comment on a webpage and copyright springs to life (in contrast to a patent, which is a government examined and granted right) or where a good faith error could be made in matching every bit of content to the rightful registered owner.

Notably, because these are statutory damages, penalties would be made available for simple failure to provide a perfect list and for each violation therein – no actual harm would have to be demonstrated for there to be an award of damages. In fact, while the plaintiff has a variety of remedies to choose from, the courts have no discretion whatsoever in awarding the damages when sought. They are entitled to $1,000 per violation or actual damages, whichever is greater.

Under the newest amendments, each day following the 30-day period that a developer fails to provide the required information constitutes a discrete violation, potentially leading to astronomical penalties for technical delays.

SUPPORT

Concept Art Association (co-sponsor)
National Association of Voice Actors (co-sponsor)
SAG-AFTRA (co-sponsor)
American Federation of Musicians, Local 7
American Society for Collective Rights Licensing
Art Directors Guild (ADG IATSE Local 800)
Arte Es Ética
ASIFA-Hollywood
California Arts Advocates
California Democratic Party
California Federation of Labor Unions, AFL-CIO
Charlotte Ann Voice-overs, LLC
Coalition of Dubbing Actors
EGAIR - European Guild for Artificial Intelligence Regulation
Ethovox INC.
Fantagraphics
Graphic Artists Guild
Heart of LA Democratic Club

Iatse Local 800
Joanne's English Club
LAVA - Latin Association of Voice Actors
Los Angeles County Democratic Party
LUCITÀ Inc
Nippon Anime and Film Culture Association
Norcal Voices
Phantom Four Films
Professional Photographers of America
Santa Monica Democratic Club
SJSU Animation & Illustration Program
Society of Illustrators of Los Angeles
soundBOX: Studio Group
TechEquity Action
Transparency Coalition.ai
Writers Guild of America West
Over 23,000 individuals

OPPOSITION

AI Salon
American Innovators Network
Bay Area Council
Bizfed - Los Angeles County
Bizfed - Central Valley
Business Software Alliance
CalBroadband
California Chamber of Commerce
California Hispanic Chambers of Commerce (CHCC)
Central Valley Business Federation
Chamber of Progress
Chamber San Mateo County
Civil Justice Association of California (CJAC)
Computer & Communications Industry Association
Consumer Technology Association
Creative Commons
Developers Alliance (UNREG)
EcomBack
Electronic Frontier Foundation
Engine Advocacy
Information Technology Industry Council
Insights Association
Internet Coalition
Library Futures

Los Angeles County Business Federation (BIZFED)
Pasadena Chamber of Commerce and Civic Association
Public Knowledge
R Street Institute
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 and Commerce Association (VICA)

RELATED LEGISLATION

Pending 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 on the Senate Floor.

Prior legislation: AB 2013 (Irwin, Ch. 817, Stats. 2024) establishes the California AI Transparency Act, which 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:

Senate Privacy, Digital Technologies, and Consumer Protection Committee (Ayes 6, Noes 2)
Assembly Floor (Ayes 45, Noes 16)
Assembly Judiciary Committee (Ayes 9, Noes 1)
Assembly Privacy and Consumer Protection Committee (Ayes 10, Noes 3)

****