Date of Hearing: March 18, 2025

Fiscal: No

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
AB 412 (Bauer-Kahan) – As Amended March 10, 2025

PROPOSED AMENDMENTS

SUBJECT: Generative artificial intelligence: training data: copyrighted materials

SYNOPSIS

The world's most profitable generative artificial intelligence (GenAI) models were trained on vast datasets scraped from the internet. Within many of these datasets are troves of copyrighted books, songs, visual arts, and other artistic works. GenAI models can mimic these works at an industrial scale, threatening to displace the writers, musicians, and artists whose creative expression enabled this technological breakthrough. This has spurred numerous lawsuits by creatives who allege that these companies misappropriated their intellectual property without consent or compensation.

Co-sponsored by SAG-AFTRA, the Concept Art Association, National Association of Voice Actors, and Authors Guild, this bill seeks to increase transparency and accountability in the training of GenAI. The bill requires certain developers of GenAI models to document and disclose any registered copyrighted works used in their training datasets. They may either make this information public or provide copyright owners with a mechanism to request a comprehensive list of their works contained in the training data. The bill also provides for a cause of action for creators to hold noncompliant developers accountable.

The bill is opposed by numerous industry associations, including California Chamber of Commerce and TechNet, as well open-internet advocates, such as Electronic Frontier Foundation. They raise several common concerns, including that the bill interferes with pending litigation and recent legislation, may be preempted, risks revealing trade secrets, and is fundamentally unworkable. Amendments described in Comment 8 substantially revise the bill to address workability concerns.

If passed by this Committee, the bill will next be heard by the Assembly Judiciary Committee.

THIS BILL:

1) Defines:

    a) "Copyrighted material" as a material registered with the United States Copyright Office.

    b) "Copyright owner" as the owner of a copyright enforceable under federal copyright laws.

    c) "Developer" as a person, partnership, or corporation that designs, codes, produces, or substantially modifies a GenAI system or model and uses it commercially in California or makes it available to Californians for use.

2) Requires developers of GenAI models or systems to do all of the following:

    a) Document any copyrighted materials, and the owner thereof, that were used to train the system or model.

    b) Make such documents publicly available, or do the following:

        i) Provide a mechanism on the developer's website allowing a copyright owner to submit a written request for a comprehensive list of their copyrighted materials used to train the system or model. The developer must document such requests.

        ii) Within 7 days of receiving a request, provide the copyright owner with all such materials. If the there are none, the developer must notify the copyright owner within 30 days of receiving the request.

    c) Retain required documentation for as long as the system or model is used commercially, plus 10 years.

3) Enables copyright owners who do not receive a timely response pursuant to 1(b)(i), above, to bring a civil action against the developer for the greater of $1,000 per violation or actual damages, and to obtain injunctive or declaratory relief, reasonable attorney's costs and fees, and any other relief the court deems appropriate.

EXISTING 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 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. (*Id.* § 504(c).) Provides for injunctive relief (*id.* § 502(a)), costs and attorney's fees (*id.* § 505), and, in certain circumstances, criminalizes willful copyright infringement for commercial advantage or private financial gain (*id.* § 504(c)).

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 by (2), above. (*Id.* § 301(b)(3).)

5) 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.)

COMMENTS:

1) **Author's statement.** The author writes:

> 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) **Copyright Act.** "[T]o promote the progress of science and useful arts" the U.S. Constitution endows Congress with the power to "secur[e] for limited times to authors and inventors the exclusive right to their writings and discoveries."[1] The federal Copyright Act protects "original works of authorship fixed in any tangible medium of expression . . . ."[2] Six exclusive rights flow from copyright ownership. These are the rights to: (1) reproduce and make copies of an original work; (2) prepare derivative works based on the original work; (3) distribute copies to the public by sale or another form of transfer, such as rental or lending; (4) publicly perform the work; (5) publicly display the work; and (6) perform sound recordings publicly through digital audio transmission.[3] An infringer is liable for actual damages and any additional profits, or for statutory damages between $750 and $30,000, as the court considers just.[4] The act specifically allows for state rights and remedies for "activities violating legal or equitable rights that are *not equivalent* to any of the exclusive rights" specified above.[5]

Copyright exists from the moment the work is created.[6] In order to sue for infringement, the owner 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.[7]

3) **Training AI.** "Artificial intelligence" refers to the mimicking of human intelligence by artificial systems, such as computers.[8] AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process: numbers, text, audio, video, or movement. AI is not fundamentally different from other computer functions; unlike other computer functions, however, AI is able to accomplish tasks that are normally performed by humans.

---

[1] U.S. Const., art. I, § 8, cl. 8.
[2] 17 U.S.C. § 102.
[3] *Id.* § 106.
[4] *Id.* § 504(c).
[5] *Id.* § 301(b)(3), emphasis added.
[6] *Copyright in General: Frequently Asked Questions*, United States Copyright Office, https://www.copyright.gov/help/faq/faq-general.html#register.
[7] *Public Catalog*, United States Copyright Office, https://cocatalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First.
[8] AB 2885 (Bauer-Kahan; Ch. 843, Stats. 2024) defined the term as "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."

AI that are trained on small, specific datasets in order to make recommendations and predictions are sometimes referred to as "predictive AI." This differentiates them from GenAI, which are trained on massive datasets in order to produce detailed text, images, audio, and video. When Netflix suggests content to a viewer, its recommendation is produced by predictive AI that is trained on the viewing habits of Netflix users.[9] When ChatGPT generates text in clear, concise paragraphs, it uses GenAI that is trained on the written contents of the internet.[10] GenAI are able to quickly create complex outputs based on user inputs.

Most modern AI tools are created through a process known as "machine learning." During machine learning, an AI is exposed to data and allowed to automatically explore its structure.[11] The process of exposing a naïve AI to data is known as "training." The algorithm that an AI develops during training is known as its "model." Training is the secret sauce of machine learning. At its core, training is an optimization problem wherein a model attempts to identify specific parameters – "weights" – that minimize the difference between predicted outcomes and actual outcomes. How an input is transformed into an output depends on the specific algorithm that is developed by a model. Once trained, AI can be used to output new, never-before-seen data. An AI's performance is directly impacted by the quality, quantity, and relevance of the data used to train it.[12]

Datasets are structured collections of data used to train AI, providing the raw material used by the model to identify patterns, make predications, and generate outputs. The intended purpose of the model dictates the types of datasets used:

- Text-based models must be trained on vast corpuses of text, including books, articles, and online discussions. Widely used text datasets include Common Crawl, which includes billions of web pages, and BooksCorpus, which consists of over 11,000 books.

- Image generators are trained on collections of images that are often labeled to help discern recurrent features such as colors and shapes. ImageNet is a dataset with over 14 million labeled images across 20,000 categories. CelebA is a dataset of celebrity faces used for face-related generative outputs.

- Audio generators require recorded speech, sounds, and music to generate realistic audio. VoxCelb is a dataset consisting of speech from thousands of individuals from different demographics. The Million Song Dataset is a freely available collection of audio features and metadata for a million music tracks.

- Video-based models must ingest clips annotated with actions, settings, and context to enable models to capture and reproduce dynamic movements across time. YouTube-8M is a dataset with millions of videos that are labeled with content categories that used to train models to recognize patterns in the videos.

---

[9] Netflix, *How Netflix's Recommendations System Works*, https://help.netflix.com/en/node/100639.
[10] OpenAI, *How ChatGPT and Our Language Models Are Developed*,https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed.
[11] IBM, *What is machine learning?,* www.ibm.com/topics/machine-learning.
[12] Rohit Sehgal, "AI Needs Data More Than Data Needs AI," *Forbes* (Oct. 5, 2023), https://www.forbes.com/sites/forbestechcouncil/2023/10/05/ai-needs-data-more-than-data-needs-ai/.

- Large and multi-modal models are often trained on extensive and varied datasets.[13]

Many of these datasets consist of information scraped from web pages, social media, and online databases, and often include copyrighted materials.

4) **AI Copyright Lawsuits**. In an effort to obtain more data, tech companies have scraped massive quantities of media to extract information for training datasets. Scaling up the size of training datasets has enabled GenAI models to internalize and map human language, as well as to achieve a level of adaptability that enables these models to produce naturalistic outputs that readily pass for a human.[14]

This has led to a rash of lawsuits from content creators – including authors, visual artists, media companies, including the *New York Times*, and the music industry, including Universal Music Group – who allege, among other things, that using their works to train AI models that can reproduce similar content constitutes copyright infringement. Nearly every major GenAI company, including OpenAI, Meta, Microsoft, Google, Anthropic, and Nvidia, has been sued.[15]

---

[13] Vijay K, "What datasets are used to train AI models," *The AiOps* (Nov. 5, 2024), https://www.theaiops.com/what-datasets-are-used-to-train-generative-ai-models/.
[14] Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson and Nico Grant, "How Tech Giants Cut Corners to Harvest Data for AI," *New York Times* (Apr. 6, 2024), https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.
[15] Kate Knibbs, "Every AI Copyright Lawsuit in the US, Visualized" *Wired* (Dec. 2024), https://www.wired.com/story/ai-copyright-case-tracker/.

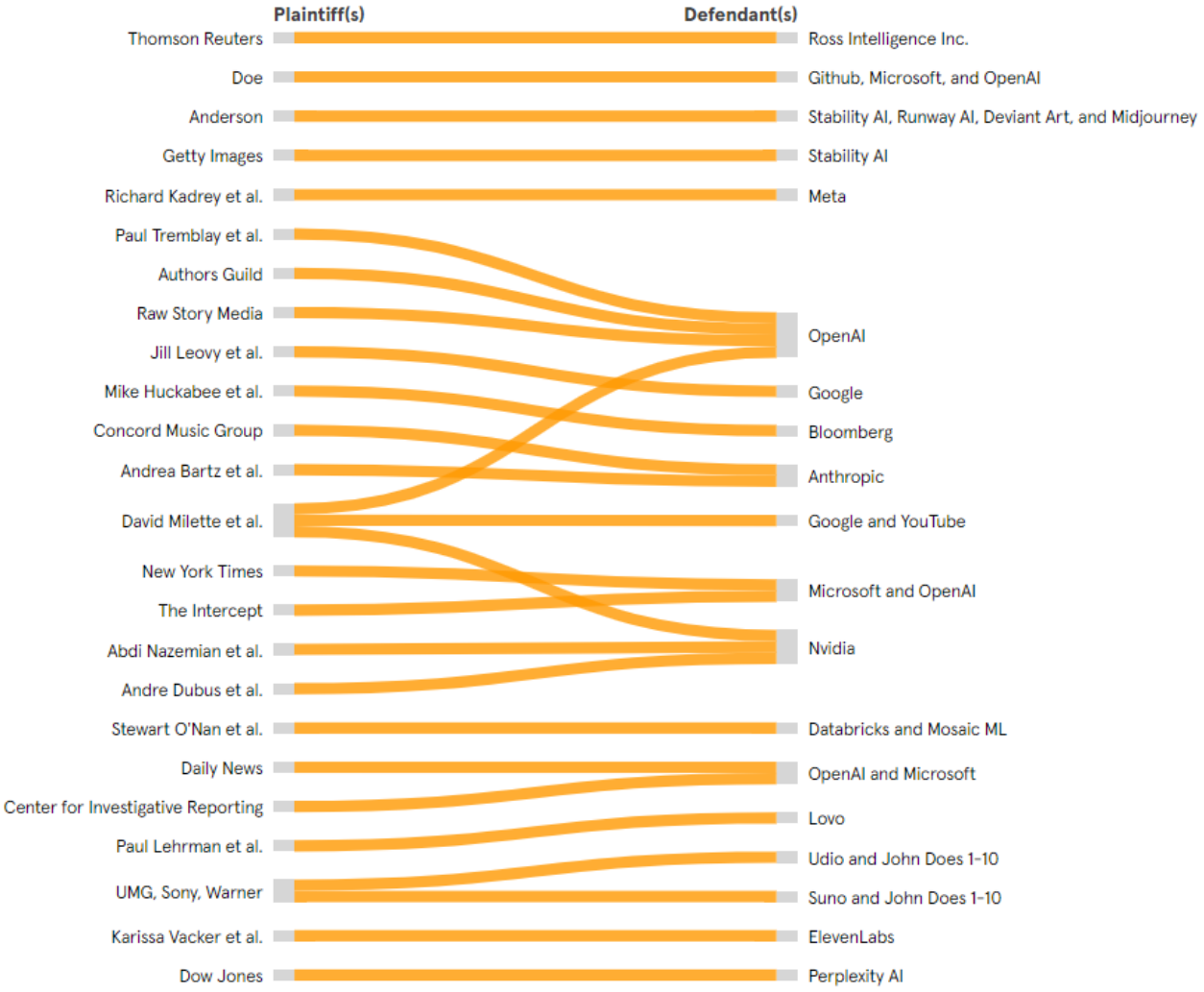

Fig. 1: Who's Suing Who? Source: *Wired*

A key issue in many of these lawsuits is whether this practice constitutes a "fair use," a defense to copyright infringement claims. A fair-use analysis considers four factors: (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.[16] As part of this analysis, courts assess whether the new use is "transformative" because it "communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge."[17] Relevant utility-expanding uses have included scanning books to create a full-text searchable database and public search function;[18] copying works into a database used

[16] 17 U.S.C § 107(1)-(4).
[17] *Authors Guild v. Google, Inc.* (2d Cir. 2015) 804 F.3d 202, 214.
[18] *Authors Guild, Inc. v. HathiTrust* (2d Cir. 2014) 755 F.3d 87, 97-98.

to detect plagiarism;[19] and displaying "thumbnail" reproductions of works to provide links to websites containing the originals.[20]

While these cases are still pending, one relevant case, *Thomson Reuters v. Ross*[21] was resolved recently in favor of the copyright holder. Thomson Reuters, the owner of Westlaw, alleged that a now-defunct AI company, Ross Intelligence, infringed its copyright by using Westlaw's headnotes – summaries used to identify cases in Westlaw's search tool – as well as a case-numbering system to train its own legal search tool. The court ruled in favor of Thomson Reuters, concluding, among other things, that Ross's direct infringement for the purpose of creating a competing tool was neither transformative nor fair.[22] It should be noted, however, that this case involved predictive, not generative, AI.

5) **Recent efforts to increase transparency in training data.** At the federal level, there are two relevant recent bills. H.R 7913, the Generative AI Copyright Disclosure Act of 2024, which was introduced by then Congressmember Adam Shiff, would have required any person who creates or alters a training dataset used to build a GenAI system to submit to the Copyright Register a notice containing a sufficiently detailed summary of any copyrights works used, as specified.

The Transparency and Responsibility for Artificial Intelligence Networks Act, introduced last year by Senator Peter Welch, would have enabled the legal or beneficial owner of an exclusive right under a copyright to request the clerk of any U.S. district court to issue a subpoena to a developer or deployer for disclosure of copies of, or records sufficient to identify with certainty, the works likely owned or controlled by the requester that were used to train a GenAI model, provided that the requester has a subjective good faith belief that the developer or deployer used some or all of one or more such copyrighted works to train the GenAI model.

Here in California, AB 2013 (Irwin, Ch. 817, Stats. 2024) requires certain developers of publicly available GenAI systems or services, by January 1, 2026, to post on their websites a high-level summary of the datasets used in the development of the system or service. The summary must include, among other categories, information about data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain.[23] Although silent as to enforcement, the bill can be enforced under the Unfair Competition Law.[24] The bill passed this Committee by an 8-1 vote.

6) **This bill.** AB 412 requires developers of GenAI models or systems to document any registered copyrighted materials that were used to train the system or model, as well as the owners of such materials. The developer then has the option of either making that documentation public, or providing a mechanism that enables a copyright owner to submit a written request for a comprehensive list of their copyrighted materials used to train the system or model. If a copyright owner submits such a request, the developer must provide the owner with a comprehensive list of such materials within seven days, unless there are no such materials, in which case the developer must respond to the copyright owner within 30 days. Developers must

---

[19] *A.V. ex rel. Vanderhye v. iParadigms, LLC* (4th Cir. 2009) 562 F.3d 630, 639.
[20] *Perfect 10, Inc. v. Amazon.com, Inc.* (9th Cir. 2007) 508 F.3d 1146, 1165; *Kelly v. Arriba Soft Corp.* (9th Cir. 2003) 336 F.3d 811, 818-19.
[21] (2/11/25) U.S. Dist. Ct N.Del. No. 1:20-cv-00613-SB.
[22] *Id.* at p. 19.
[23] Civ. Code § 3111.
[24] Bus. & Prof. Code § 17200 et seq.

retain required documentation for as long as the system or model is used commercially, plus 10 years.

The bill also enables copyright owners who do not receive a timely response to a request to bring a civil action against the developer for the greater of $1,000 per violation or actual damages, and to obtain injunctive or declaratory relief, reasonable attorney's costs and fees, and any other relief the court deems appropriate.

A coalition of supporters writes: "GenAI developers frequently use copyrighted materials to train new systems without crediting or compensating the owners of those materials. In order for a copyright owner to exercise their rights with respect to a copyrighted material, the owner must first know how the material is being used. At present, copyright owners have no way of knowing whether a given copyrighted material was used to train a GenAI system or model."

7) **Opposition concerns.** The bill is opposed by numerous industry associations, including California Chamber of Commerce and TechNet, as well open-internet advocates, such as Electronic Frontier Foundation. They raise several common concerns, including the following:

The bill is premature because AB 2013 has not become operative. As described above, AB 2013 requires GenAI developers to disclose high-level summaries of various categories of information regarding training datasets, including whether copyrighted materials were used to train GenAI models. This requirement becomes operative January 1, 2026. But it can be argued that this bill supplements that bill by giving copyright holders a complete understanding of how many of their works were used to train GenAI models or systems.

The bill could reveal trade secrets. Opponents argue that identifying all copyrighted materials within the training datasets for proprietary models could, in effect, reveal trade secrets. The industry coalition writes: "How a model is trained and on what data is an incredibly valuable piece of information and is what makes AI companies worthy of significant investment." But it can be argued that this intellectual property concern ignores the underlying intellectual property concern – whether training on copyrighted material is fair use – that is still being litigated.

The bill will lead to excessive litigation. Several opponents argue the bill will open litigation floodgates. Recent amendments to the bill in print mitigate these concerns to a degree, as they clarify that the bill is limited to registered copyright materials.

The bill is preempted by federal copyright law. Silicon Valley Law Group writes, "AB 412 treads into territory that provokes federal preemption. Copyright law is reserved for the federal government." But as mentioned above, the Copyright Act specifically allows for state laws that are "not equivalent" to the six exclusive rights provided under the Act, none of which involve transparency. This issue is the province of the Judiciary Committee, which awaits this bill if it passes out of this Committee.

The bill is unworkable. Industry opponents write: "At its core AB 412 is manifestly impossible." They continue:

> AB 412 especially has a negative impact on using publicly sourced data to train GenAI as it would force developers to pick through such data sourced through the web to see if there are potentially registered copyrighted works and match them with copyright owners. While there are efforts underway to include metadata in copyrighted works to show provenance,

this work is just starting. The amount of data and the number of datasets used to train AI models cannot be understated. In order to be able to provide a list as required by this bill, a company would have had to have categorized potentially trillions of data points prior to training. To do that accurately for copyright holders, the database of copyrights would need to be machine-readable and searchable, which it is not at this point in time.

This type of cataloging and matching of data to registered copyrights would be an incredibly expensive and massive undertaking for even the largest company, but would be especially severe for smaller companies who are trying to compete and simply don't have the resources to employ staff or engineer a process to comply with this bill. For these reasons alone, providing a list upon written request from a copyright owner within 7 days is simply impossible.

To mitigate compliance burdens and make the bill more workable, the author has agreed to amend the bill, as described below.

8) **Amendments to address compliance challenges.** Following a series of copyright lawsuits between 2007 and 2009, YouTube developed ContentID to enable copyright holders to limit infringements on the platform. Under this program, exclusive owners of copyrights are able to upload copyrighted content into a database. Newly uploaded videos are automatically scanned against the database to determine if there is a match. If so, content owners are given the choice of blocking, tracking, or monetizing the infringing content.[25]

While a system such as ContentID would likely work for most developers, it poses challenges for creators who may be understandably hesitant to provide their content to a developer who may train on that very content. Is there a middle ground between the bill in print and an approach such as this?

A technique for finding similar files in large repositories known as "approximate fingerprinting" offers a potentially elegant solution. "Approximate fingerprints provide a compact representation of a file such that, with high probability, the fingerprints of two similar files are similar (but not necessarily equal), and the fingerprints of two non-similar files are different."[26] Approximate fingerprints provide an efficient, accurate way of finding matching content in large datasets without having to upload an entire file to perform the search. Approximate fingerprints have been used for detecting code plagiarism, network intrusion and malware, spam emails, and image classification tasks.[27]

University of Chicago Professor of Computer Science Ben Zhao, in support of the bill, writes the following with respect to using approximate fingerprint principles in AB 412:

---

[25] "Using Content ID," YouTube Operations Guide, https://support.google.com/youtube/answer/3244015?hl=en#:~:text=Content%20ID%20is%20YouTube%27s%20automated%2C%20scalable%20system%20that,ID%20to%20copyright%20owners%20who%20meet%20specific%20criteria.

[26] Udi Manber, "Finding similar files in a large file system" In Proc. of USENIX Winter Technical Conference, volume 94, p. 2, 1994, https://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/manber94finding.pdf.

[27] Li et al, "Blacklight: Scalable Defense for Neural Networks against Quer-Based Black-Box Attacks," Proc. of USENIX Security 2022 video presentation and link to paper pdf, p. 2122, https://www.usenix.org/system/files/sec22-li-huiying.pdf.

Feasibility and Understanding of Approximate Fingerprints as a signature tool.

Well understood. Approximate fingerprints as a general concept has been known for decades. They were first proposed in 1993 by researchers at then Yahoo Research,[28] and deployed as a tool to identify similar documents. Since then, there have been numerous research proposals suggesting the use of similar algorithms in the database community, the web community, information retrieval and security communities, including a paper published by Google's co-founder Sergey Brin as a Stanford PhD student.[29] These tools are well understood, and multiple implementations exist in the public domain, using bloom filters, locality sensitive hashes, and randomized hash sampling methods. The broad range of implementations and possible parameters give additional flexibility and control over the tradeoff between more accuracy (eliminating false positives or false negatives) and robustness (allowing fingerprints to identify the same document or files despite small tweaks or errors across versions).

Fast. It is important to note that most of these fingerprints are quite fast to compute and only incur a one-time overhead on the training data. For example, variants described by the Manber 1993[30] paper and the Zhao 2003[31] paper require only a single (lightweight) pass through the data to generate all fingerprints. String hashing itself is a very low cost process that is easily parallelized. Consider that any training data goes through multiple rounds of data curation and preprocessing before being used as training data for an AI model, computing approximate fingerprints would be a ONE-TIME cost that is substantially faster/lower cost than any curation process. The overhead imposed on AI companies would be negligible and difficult to measure.

Generalizable. Approximate fingerprints can be easily generalized across different data types. Existing literature already describes its use for text documents and images (Blacklight, Li 2022).[32] It would not be difficult to generalize it to audio files, even music recordings, and other new modalities in the future (e.g. 3-D models, short videos). Thus as a general mechanism, it is likely to be applicable even as different data types emerge in the years ahead.

The proposed amendments to AB 412 broadly align with these principles, making the policy objectives of the bill low-cost and scalable across content modalities.[33]

The amendments referenced by Professor Zhao, and agreed to by the author, would do the following:

- Limit the bill's application to developers of GenAI models, not systems. If models are the brains of GenAI applications, systems are bodies that allow GenAI to perform useful

---

[28] "Finding similar files in a large file system," *supra*.
[29] Brin et al, "Copy detection mechanisms for digital documents," In Proceedings of ACM SIGMOD 1995 https://dl.acm.org/doi/10.1145/223784.223855.
[30] "Finding similar files in a large file system," *supra*.
[31] "Approximate object location and spam filtering on peer-to-peer systems," In Proceedings of the ACM/IFIP/USENIX 2003 International Conference on Middleware, https://disco.ethz.ch/courses/ws0506/seminar/papers/peer_spam_filtering.pdf.
[32] "Blacklight: Scalable Defense for Neural Networks against Quer-Based Black-Box Attacks," *supra.*
[33] Bold and italics added.

work. The bill is intended to apply at the model training level rather than to specific downstream applications of models at the system level, which would create significant compliance challenges for system developers lacking access to original training data.

- Instead of requiring developers to proactively identify all copyrighted content in their datasets, developers would be required to make "reasonable efforts" to do so.

- Instead of enabling copyright owners to request a comprehensive list of copyrighted materials, copyright owners would be able to provide developers with an approximate content fingerprint of their registered copyrighted materials, which the developer would then use to search for matches in the datasets. Developers must provide information sufficient for copyright owners to generate approximate fingerprints of their materials.

- Exempt developers that make all of the data used to train their GenAI models publicly available at no cost.

The amendments are as follows:

3115. For the purposes of this title, the following definitions apply:

(a) "Approximate content fingerprint" means an abstract representation of digital content that encodes distinctive features of the content and that is all of the following:

(1) Distinct to the digital content being represented.

(2) Robust to minor variations in the original digital content.

(3) Incapable of being used to reconstruct the original digital content.

(4) Capable of being used to readily identify digital content in a dataset.

(~~a~~***b***) "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~~***c***) "Copyrighted material" means a material registered with the United States Copyright Office pursuant to Title 17 of the United States Code, Public Law 94-553 (17 U.S.C. Sec. 101 et seq.).

(~~c~~***d***) "Copyright owner" means the owner of a copyright enforceable under the copyright laws of the United States pursuant to Title 17 of the United States Code, Public Law 94-553 (17 U.S.C. Sec. 101 et seq.).

(~~d~~***e***) "Developer" means a person, partnership, or corporation that designs, codes, produces, or substantially modifies a GenAI ~~system or~~ model and that does either of the following:

(1) Uses the GenAI ~~system or~~ model commercially in California.

(2) Makes the GenAI ~~system or~~ model available to Californians for use.

(~~e~~***f***) "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.

3116. ~~(a)~~ A developer of a GenAI ~~system or~~ model shall do all of the following:

(a) (1) ~~(A)~~ Document any copyrighted materials ***that the developer knows were*** used to train the GenAI ~~system or~~ model.

~~(B) For~~ ***(2) Make reasonable efforts to identify and document any other copyrighted materials that were used to train the GenAI model.***

(3) Document the copyright owner of each copyrighted material documented pursuant to this subdivision~~, document the copyright owner of the copyrighted material~~.

~~(2) (A~~***(b) Make available information on the developer's internet website sufficient to enable a copyright owner to generate an approximate content fingerprint for any copyrighted material that might have been used to train the developer's GenAI model.***

(c) Make available a mechanism on the developer's internet website allowing a copyright owner to submit a written request pursuant to ***subdivision (a) of*** Section 3117.

(~~B~~***d***) Document any requests received pursuant to ***subdivision (a) of*** Section 3117.

(~~3~~***e***) Retain any documentation required by this section for as long as the ***developer uses the*** GenAI ~~system or~~ model ~~is used~~ commercially ***in California or makes the GenAI model available to Californians for use***, plus 10 years.

~~(b) Paragraph (2) of subdivision (a) does not apply to a developer that makes the documentation required under paragraph (1) of subdivision (a) publicly available at no cost to users of the developer's internet website.~~

3117. (a) ***(1) A copyright owner may request information about a developer's use of copyrighted materials held by the copyright owner by providing the developer with all of the following:***

(A) Proof of the copyright owner's identity.

(B) The copyright registration numbers of each copyrighted material held by the copyright owner.

(C) An approximate content fingerprint for each copyrighted material.

(2) A developer's collection, use, retention, and sharing of information from a copyright owner pursuant to this subdivision 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.

(b) (1) Within seven days of receiving a ~~written~~ request from a copyright owner ~~of a copyrighted material used~~***pursuant*** to ~~train~~***subdivision (***~~a GenAI system or model.~~***),*** a developer shall provide the copyright owner with a ~~comprehensive~~***complete*** list of copyrighted materials ***held by the copyright owner that were*** used to train the GenAI ~~system or~~ model ~~for which the copyright owner holds the copyright~~.

(2) Each day after the seven-day period described in paragraph (1) that a developer fails to provide a copyright owner with a list of copyrighted materials ~~pursuant to this subdivision~~ constitutes a discrete violation of this title.

~~(b) Within 30 days of receiving a written request from a copyright owner whose copyrighted materials were not used to train a GenAI system or model, a developer shall notify the copyright owner that no copyrighted materials for which the copyright owner holds the copyright were used to train the GenAI system or model.~~

~~(c) This section does not apply to a developer that makes the documentation required under paragraph (1) of subdivision (a) of Section 3116 publicly available at no cost to users of the developer's internet website.~~

3118. A copyright owner that is not provided with a list of copyrighted materials ~~or notified by a developer~~ as required by this title may bring a civil action against the developer for any of the following:

(a) One thousand dollars ($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.

3119. This title does not apply to a developer that makes all of the data used to train the developer's GenAI model publicly available at no cost to users of the developer's internet website.

ARGUMENTS IN SUPPORT: Writing about the bill in print, California Civil Liberties Advocacy, in support, states:

AB 412 provides a reasonable and necessary safeguard by requiring AI developers to document copyrighted materials used in AI training and provide copyright owners with access to this information upon request. This bill promotes fairness, upholds intellectual property rights, and ensures that artists, writers, and content creators receive the protections they are entitled to under existing copyright laws.

Additionally, AB 412 offers a balanced approach that encourages responsible AI innovation while preventing the misuse of copyrighted works. By allowing developers to publicly disclose training data to avoid individual disclosure requests, the bill maintains flexibility for AI companies while ensuring accountability.

The passage of AB 412 will:

- Enhance transparency in AI development and protect original creators.
- Encourage ethical AI practices by preventing unauthorized use of copyrighted materials.
- Provide legal recourse for copyright owners whose works are used without proper disclosure.
- Maintain a fair balance between fostering AI innovation and protecting intellectual property rights.

ARGUMENTS IN OPPOSITION: A coalition of industry opponents, led by Chamber of Commerce, writes of the bill in print:

> [W]e have serious concerns about **AB 412** as its burdensome requirements disadvantage smaller AI companies and startups and undermine recently passed legislation. Given the complexity of the subject matter involved and the early nature of the hearing, we are still evaluating the bill and may have other issues but have attempted to identify as many as possible in this letter. However, we ultimately feel that this bill is neither necessary due to the very 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, and interference with pending litigation . . .

In opposition to the bill in print, Electronic Frontier Foundation states:

> EFF respectfully writes to express our grave concerns about A.B. 412, a bill that would create a new state-level requirement to discern and disclose information about uses of in-copyright works. A.B. 412 attempts an unlawful end-run around federal copyright law and imposes an impossible new regulatory regime that would cause devastating collateral damage for research and innovation. We also believe it would ultimately do little to help creators receive just compensation.

In opposition to the bill in print, Silicon Valley Leadership Group states:

> [C]ompliance with the strictures of AB 412 would be unworkable because any training completed by use of material on the internet would require that developers catalog every piece of potentially copyrighted material in any format and then attempt to locate it in the database. For the open web, this is not practicable. As a result, AB 412 would have the effect of essentially prohibiting model training on the open web. Developers would have access to less robust and diverse datasets, resulting in a greater risk for models that are less representative of diverse communities and more prone to bias.

REGISTERED SUPPORT / OPPOSITION:

Support

Writers Guild of America West (co-sponsor)
California Civil Liberties Advocacy
Concept Art Association

Music Artists Coalition
National Association of Voice Actors
National Writers Union
Romance Writers of America, INC.
Sag-aftra
Songwriters of North America
41 Individuals

Opposition

ACT / the App Association
California Chamber of Commerce
Civil Justice Association of California
Computer and Communications Industry Association
Creative Commons
Electronic Frontier Foundation
Engine Advocacy
Entertainment Software Association
Insights Association
Library Futures
Public Knowledge
R Street Institute
Re:create
Silicon Valley Leadership Group
Technet

Analysis Prepared by: Josh Tosney / P. & C.P. / (916) 319-2200