July 8, 2025

TO: Members, Senate Judiciary Committee

SUBJECT: AB 412 (BAUER-KAHAN) GENERATIVE ARTIFICIAL INTELLIGENCE: TRAINING DATA: COPYRIGHTED MATERIALS
OPPOSE – AS AMENDED MAY 7, 2025
SCHEDULED FOR HEARING – JULY 15, 2025

The undersigned organizations respectfully **OPPOSE AB 412 (Bauer-Kahan)** as amended May 7, 2025, which requires a developer that makes a generative artificial intelligence (GenAI) system or model available to Californians for use to, among other things, document both the copyrighted materials used to train the system model and the copyright owner of that material. Under U.S. law, copyright exists at the moment an original work of authorship is fixed in tangible form. There is no requirement of registration or formal application—the copyright exists automatically. There is also no requirement that a copyright be associated with its producer—even anonymous works are copyrighted at fixation. As a result, essentially all training data is "subject to copyright protection." This makes compliance with this bill inherently difficult, if not impossible.

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** 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, among other things.

AB 412 requires disclosure of trade secrets and intellectual property, undermining the careful balance of recently enacted legislation which addressed transparency in data used to train GenAI and undercutting California's status as an AI leader

We are concerned that **AB 412** lacks necessary protections for intellectual property and fails to include any other reasonable limitations on the scope of its transparency which will ultimately require the disclosure of trade secrets and other sensitive material. 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. Requiring companies to not only disclose this information but list it publicly jeopardizes this value and risks undercutting California's status as a global leader on AI.

These risks seem particularly unnecessary given the recent passage of AB 2013 (Irwin), which the Legislature passed, and the Governor signed less than six months ago. In contrast to **AB 412**, AB 2013 mandated that generative AI companies disclose information regarding the data used to train the generative AI system but also balanced the interests of transparency and confidentiality by requiring companies to

provide summaries of the sources and descriptions of datasets, among other categories, used to train the generative AI system as well as whether the datasets included data protected by copyright, trademark, or patent. (See Civil Code Sec. 3111.) AB 2013 has not even gone into full effect as it allowed companies to gather the necessary information on currently existing systems and make the required disclosure by January 1, 2026.

AB 412 appears modeled upon that bill yet also deviates from critical aspects and amendments that were agreed upon in the Assembly Privacy & Consumer Protection Committee that were intended to resolve these very concerns. Worse, it creates such risks unnecessarily, given that 2013 already addresses the need for transparency in GenAI's training data. As such, we suggest waiting until there has been sufficient time to judge the efficacy of AB 2013 before enacting sweeping, new AI transparency legislation, particularly given that AB 2013 already applies to GenAI systems and services. If there is a problem with AB 2013, or if there is a problem with how the courts resolve copyright issues around GenAI training data (see below), the Legislature should revisit the issue at a later date when it has more information.

At its core AB 412 is manifestly impossible and the impact that it will have on training AI models should be taken seriously

Documenting the copyright owner of every piece of content that might have been used is simply not feasible when everything is copyrighted from the start. Virtually everything that is not in the public domain effectively becomes copyright material as soon as it is placed into fixed form—including something as simple as a 140-character social media "tweet". Currently central questions exist over what materials are copyrightable, and courts continue to entertain questions of what types of works are entitled to copyright protection on a case-by-case basis. In fact, in 2023, the Copyright Office only issued 441,526 registrations[1]. A developer cannot be expected to make billions of judgments about whether material found on the web qualifies as copyrighted works.

Moreover, whether or not the bill is limited to registered copyrights, there is no central database of authors associated with every piece of work publicly available on the web. Questions of authorship plague the copyright system today – such as in the case of Orphan Works or works that may include contributions from multiple authors. It is simply incomprehensible, let alone unreasonable, to consider that an AI developer would be able to understand the copyright owner associated with billions of works found on the web.

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 even within 30 days is simply impossible.

Amendments do not fully resolve the technical infeasibility of AB 412

As noted above, on a fundamental level, cataloging copyrighted works used in AI training and identifying their owners is impractical at best, and impossible in all likelihood. This type of burdensome requirement will directly affect the ability of our members to train large language models and eventually impact the quality of those models as a result.

[1] U.S. Copyright Office • FY 2024 Facts at a Glance

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, making this bill vastly impractical, if not simply infeasible, 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 only further complicated for works created prior to 1978 given the fact that the U.S. Copyright Office only permits searches of registration records for works from this period to be conducted in-person, in their Washington, DC office.

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.

Relatedly, take the proponents' comparison of the proposed fingerprinting mechanism to the ContentID system. Such a comparison to demonstrate the feasibility of the bill paints an inaccurate picture of how this would translate in practice. The body of content monitored by ContentID is vastly smaller than the content needed for model training, as is the number of rights holders who are eligible to participate. Expanding it to the volume of data needed for model training and to all copyright holders will necessarily create a host of issues, particularly given that ContentID is a system designed to deal with copyrighted works that are duplicated and posted online – an activity that, if it has not been authorized by the only rights holder, is likely to amount to copyright infringement. This is quite different to a work appearing in an initial set of AI training data – an activity that many are currently arguing in court is a protected fair use.

Furthermore, we note that these amendments only require the rights holder to prove their identity in order to gain untethered access to training data that they may not have the rights to. These provisions, without further clarity, would allow malicious actors the ability to obtain sensitive, proprietary, and private training data they previously had no access to before the most recent amendments.

Ultimately, questions of authorship plague the copyright system today and under **AB 412,** model developers would be required to make billions of judgments about whether data used to train their model is the subject of a registered copyright based off a request by an individual who is not even obligated to provide them with sufficient documentation to verify their identity as the rights owner of the material in question. Making that many decisions is statistically certain to result in inadvertent errors, which this bill would punish with lawsuits and statutory penalties. Needless to say, these requirements and the penalties that would invariably result would hit smaller developers particularly hard because they do not have the resources that larger market leaders do.

AB 412 interferes with pending litigation and potentially contradicts recent decisions

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 are currently 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[2]. 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 appear to suggest that training on this data is fair use (*see e.g., Bartz v. Anthropic* (N.D. Cal., June 23, 2025) where the court recently granted summary judgment in favor of Anthropic on the fair use defense for training its language model ("Claude") with copyrighted books). At the same time, insofar as compliance with AB 412 may require a business to maintain or reconstruct a detailed repository of works for the purpose of disclosing whether a particular author's work was used to train and AI model, compliance with AB 412 could create additional legal exposure for businesses if maintenance of a "library" of works is not equally upheld to be a fair use.

"Rights owners" may not in fact be the actual copyright owners, creating potential vulnerabilities under AB 412

First, as recently amended, the bill replaces references to "copyright owners" with "rights owners". Replacing the word *copyright* with *rights* does not change the fact that the bill is still about Copyright law, and the copyright rights of copyright owners, raising preemption concerns (see below).

Recently, in a unanimous decision, the U.S. Supreme Court held that the registration of a copyright is a prerequisite to filing a copyright lawsuit in *Fourth Estate Public Benefit Corp. v. Wall-Street.com, LLC* (2019) 139 S. Ct. 881. It is unclear if someone who owns a copyright due to creation, but has not registered it, qualifies as a copyright owner. This bill exacerbates the problem by stating that a "rights owner" is "the owner of a copyright enforceable under the copyright laws of the United States." **AB 412** does not impose any requirement to prove copyright ownership or impose a penalty for false representations. Arguably, in failing to do so, the bill upsets established federal copyright law which requires the copyright holder to demonstrate ownership of a valid copyright in a work and provide evidence of unauthorized use by another.

Notably, under Section 3117 of the bill, a developer has to provide the rights owner with a comprehensive list of "covered materials" (i.e. copyrighted materials) used to train the GenAI system or model for which the copyright owner holds the copyright within seven days of receiving a written request from a copyright owner. And because companies are likely to comply with requests rather than run the risk of an escalating penalty, malicious actors could easily use the Section 3117 mechanism to extract information from the training data that they have no rights to, potentially gaining access to trade secret data or personal information of others.

Finally, it is also worth noting here that the recent amendment requiring the provision of sufficient documentation establishing the rights owners' identity (Proposed Section 3117(d)) is not the same thing as requiring documentation establishing that the rights owner owns the rights to a copyright.

AB 412's statutory penalties are unjustified and overly broad

It is unclear to us why **AB 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

[2] *See e.g.*, 'The New York Times' takes OpenAI to court. ChatGPT's future could be on the line : NPR; Every AI Copyright Lawsuit in the US, Visualized | WIRED; New York Times sues OpenAI, Microsoft for copyright infringement | CBC News

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.

Furthermore, it's not clear what constitutes a violation and whether inadvertent errors or omissions would constitute separate violations. If that were to be the case, it would only underscore how difficult and costly compliance would be for companies to go through the effort of sorting through voluminous datasets and still be subject to distinct civil penalties for inadvertent errors.

Federal Copyright law preemption concerns

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.

For all of the aforementioned concerns, including that the bill is simply inoperable, inconsistent with and unnecessary in light of AB 2013 (Irwin), would interfere with pending litigation, and would undercut California's position as a global leader of AI, we must unfortunately **OPPOSE AB 412**.

Sincerely,

Ronak Daylami
Policy Advocate
on behalf of

CalBroadband, Amanda Gualderama
California Chamber of Commerce, Ronak Daylami
Civil Justice Association of California, Lusine Chinkezian
Computer & Communications Industry Association, Aodhan Downey
Insights Association, Howard Fienberg
Internet Coalition, Tammy Cota
Software Information Industry Association, Paul Lekas
TechCA, Courtney Jensen
TechNet, Jose Torres

cc: Legislative Affairs, Office of the Governor
Consultant, Senate Judiciary Committee
Slater Sharp, Office of Assemblymember Bauer-Kahan
Morgan Branch, Consultant, Senate Republican Caucus

RD:ldl