CALIFORNIA LEGISLATURE

STATE CAPITOL
SACRAMENTO, CALIFORNIA
95814

Artificial Intelligence and Copyright

A Joint Informational Hearing of the Senate Judiciary Committee and the Assembly Privacy and Consumer Protection Committee

Monday, December 8, 2025
10 a.m.
Paul Brest Hall, Stanford University
555 Salvatierra Walk, Stanford

Background Paper
By the Staff of the Assembly Privacy and Consumer Protection Committee and the Senate Judiciary Committee

---

I. Introduction

The use of copyrighted works in training artificial intelligence (AI) models has sparked significant legal and ethical debate. Generative AI (GenAI), in particular, can produce text, images, video, or audio that emulates the expressive works used to train it. Key concerns include whether training on copyrighted material without permission constitutes fair use or copyright infringement, particularly when models can sometimes reproduce portions of their training data. Content creators argue they should be compensated when their work is used to train commercial AI systems, while some worry that AI-generated content could compete with and displace original creative works.

There are also questions about transparency – many users and creators are unaware of which copyrighted materials were included in training datasets. Over 50 lawsuits from authors, artists, and publishers are currently testing these issues in court, seeking to establish clearer legal boundaries around AI training practices. The debate balances innovation and the advancement of AI technology against intellectual property rights and the economic interests of content creators. A number of novel approaches to handling these issues have been explored internationally, including in the European Union, Japan, and Australia, and other approaches have been discussed and proposed by leading academics.

This paper is intended to provide a brief overview of the relevant considerations in anticipation of a joint hearing of the Senate Judiciary Committee and the Assembly Privacy and Consumer Protection Committee, entitled "AI and Copyright."

II. Generative AI and Training Data

"Artificial intelligence" refers to the mimicking of human intelligence by artificial systems, such as computers. 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 other data. AB 2885 (Bauer-Kahan & Umberg; Ch. 843, Stats. 2024) defined AI 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." GenAI is a subset of AI that can produce outputs that closely resemble human-created content. AB 2013 (Irwin, Ch. 817, Stats. 2024) defined GenAI as "artificial intelligence that can generate derived synthetic content, such as text, images, video, and audio, that emulates the structure and characteristics of the artificial intelligence's training data."

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

Staggering quantities of data are required to train the most advanced models. For example, GPT-4 – 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] Because the model does not directly store its training data, but rather encodes abstract patterns gleaned from the data, the model itself can fit on a thumb drive.

---

[1] IBM, What is generative AI?, *available at* https://www.ibm.com/think/topics/generative-ai; IBM, What is machine learning?, *available at* www.ibm.com/topics/machine-learning.

[2] Schreiner, *GPT-4 architecture, datasets, costs and more leaked*, The Decoder (Jul. 11, 2023), *available at* https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/; Begum, *OpenAI Releases GPT-4: A Smarter and Faster AI-Language Model with 'Human-level Performance*,*'* Vocal Media (2023), *available at* https://vocal.media/01/open-ai-releases-gpt-4-a-smarter-and-faster-ai-language-model-with-human-level-performance.

[3] Ludvigsen, *The carbon footprint of GPT-4,* Medium (Jul. 18, 2023), *available at* https://medium.com/data-science/the-carbon-footprint-of-gpt-4-d6c676eb21ae.

Quality data is essential for ensuring that a model generates naturalistic outputs. Copyrighted materials are often among the highest quality content available. Developers have assembled massive datasets, often without obtaining consent or providing compensation to the creators of the expressive content in those datasets. Automated web crawlers are used to "scrape" content from the internet and compile it into sources such as Common Crawl, which includes over 300 billion web pages.[4] The content is downloaded, copied, filtered, scrubbed of attribution, and tokenized for training.

Although GenAI models learn patterns rather than directly storing data, models can sometimes absorb distinctive or duplicated content so deeply that they can summarize the content, mimic an artist's unique style, or even reproduce the content verbatim – a phenomenon researchers describe as "memorization."[5] Such outputs can directly compete with the original work, potentially substituting for it in the marketplace. Compounding the issue, there does not appear to be a practical way to remove specific works from a trained model, as the model's "knowledge" of that work may be distributed across billions of parameters.[6]

III. Impact of Generative AI on California's Creative Economy

a. California's Creative Economy

California's creative industries are key contributors to the state's economy. According to the Governor's Office of Business and Economic Development (GO-Biz), "California is renowned for its long and storied history in film and television production as the home of Hollywood and many of the world's largest studios. Major music and sound recording companies, publishers, design schools, and performing arts venues are found throughout the state, reflecting California's prominent role in the creative industries.[7]"

[4] Common Crawl, *available at* https://commoncrawl.org/.
[5] Freeman, *Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the* New York Times v. OpenAI *2023 Lawsuit*, arXiv (Dec. 2024), *available at* https://arxiv.org/html/2412.06370v1.
[6] Cooper et al, *Machine Unlearning Doesn't Do What you Think: Lessons for Generative AI Policy and Research* arXiv (Oct. 31, 2025), https://arxiv.org/abs/2412.06966.
[7] Governor's Office of Business and Economic Development, Creative Economy*, available at* https://business.ca.gov/industries/creative-economy/.

The Otis College 2025 Report on the Creative Economy estimates that California's creative economy employs more than 760,000 workers with an average salary of $191,000 per year, accounting for roughly 5% of the state's workforce.[8] The 2023 version of this report estimated that California's creative economy was responsible for 14.9% of the state's $3.4 trillion economy in 2021, generating over $194.1 billion in federal, state, and municipal taxes.[9]

Table 2: How Jobs are Divided Across the State's Creative Sectors

Creative Sector	Contribution to Creative Economy Jobs
New Media	31%
Film, TV, Sound	17%
Fashion	11%
Creative Goods and Products: Design & Manufa	11%
Advertising	8%
Fine Arts: Performance and Institutions	7%
Architecture and Related Services	6%
Traditional Media	4%
Managers, Independent Artists, Performers	4%

Source: Quarterly Census of Employment and Wages

Source: https://www.otis.edu/about/initiatives/documents/25-063-CreativeEconomy_Report4_250325.pdf

While the adoption of GenAI is opening up new possibilities in music, film, publishing and other creative fields, the technology's impact on California's creative economy and individual creators remains uncertain and unevenly distributed. To date, few studies have assessed how AI will affect jobs and incomes for writers, actors, musicians, designers and other creators.

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

> In an unchanged regulatory framework, creators will actually suffer losses on two fronts: the loss of revenues due to the unauthorised use of their works by Gen AI models without remuneration; and replacement of their traditional revenue streams due to the substitution effect of AI-generated outputs, competing against human-made works.[10]

[8] Otis College of Art and Design, Otis College Update on the Creative Economy, (Mar. 27, 2025), *available at* https://www.otis.edu/about/initiatives/documents/25-063-CreativeEconomy_Report4_250325.pdf.
[9] Otis College of Art and Design, Otis College Report on the Creative Economy, (2023), *available at* https://www.televisionacademy.com/files/assets/Downloads/2023-otis-report-on-the-creative-economy-v1.pdf.
[10] PMP Strategy/CISAC, Study on the economic impact of Generative AI in the Music and Audiovisual industries, (Nov. 2024), *available at* https://www.cisac.org/services/reports-and-research/cisacpmp-strategy-ai-study.

Meanwhile, the market for AI-generated music is projected to surge in the coming years. The CISAC study projected that the market for AI-generated music would grow to €64 billion ($74.3 billion) in 2028, with revenues for AI music platforms alone rising to about €4 billion ($4.64 billion). GenAI is forecast to claim roughly 20% of streaming music revenues by 2028, and AI-generated music has begun to regularly appear on the Billboard and Spotify charts.[11] One *Guardian* report found an entirely AI-created band – the Velvet Sundown – had amassed over one million Spotify streams.[12]

In January 2024, consulting firm CVL Economics surveyed 300 entertainment industry leaders to assess the impact of GenAI on these industries, in collaboration with the Animation Guild, the Concept Art Association, the Human Artistry Campaign, and the National Cartoonists Society Foundation.[13] They found that 203,800 entertainment jobs would likely be disrupted in the United States by 2026, including 62,000 jobs in California. In their survey, creative leaders identified certain roles – including 3D modeling, animation, and concept art – as being especially vulnerable to AI. 75% of respondents said GenAI tools had already led to the elimination or consolidation of jobs in their companies, while only about 26% felt their workforces were fully prepared for AI integration.

b. Artists' perspectives

Although some artists have found ways to use GenAI to enhance their work, overall it appears that artists' outlook on AI is generally negative. In early 2025, Mozilla Foundation and the Berggruen Institute brought together 91 people across five gatherings in Los Angeles to interrogate the growing role of AI in the creation of art. The study found that "creative labor is increasingly precarious and devalued. AI compounds gig instability, drives down value, and threatens sustainability of artistic careers." The majority of participants – over 63% of creatives and 75% of technologists – did not feel represented by any institution in the fight for creative rights in tech. According to one participant:

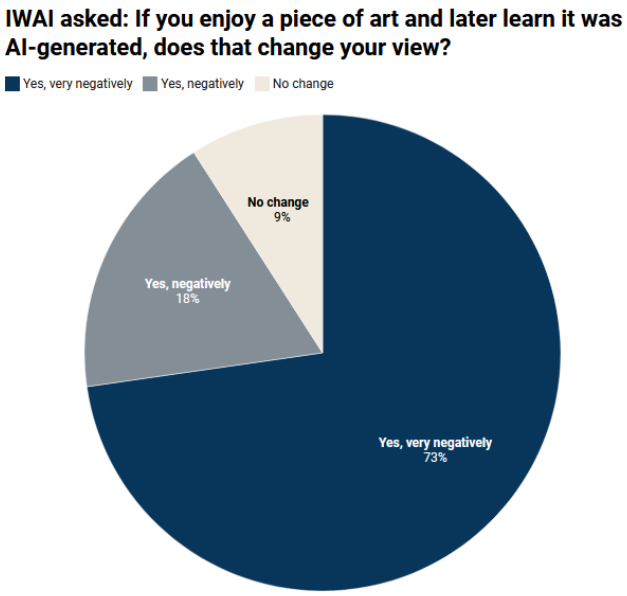
> "It's become a lot less sustainable for artists to be able to pursue an artistic career if they do not already come from money. The big fear with AI right now is that this is another way to devalue the work of artists… because it is 'easy to produce', because 'all you got to do is put a bunch of prompts in and generate something out of that. It means that your value as an artist is going to go down, and it means that it's going to be a lot less sustainable for you to pursue it as a career… Do you know how many gigs I have right now just to be able to sustain my living right now?...

---

[11] Down, *AI slop tops Billboard and Spotify charts as synthetic music spreads,* The Guardian (Nov. 13, 2025), *available at* https://www.theguardian.com/technology/2025/nov/13/ai-music-spotify-billboard-charts.
[12] Bakare, *An AI-generated band got 1m plays on Spotify. Now music insiders say listeners should be warned,* The Guardian (Jul. 14, 2025), *available at* https://www.theguardian.com/technology/2025/jul/14/an-ai-generated-band-got-1m-plays-on-spotify-now-music-insiders-say-listeners-should-be-warned.
[13] CVL ECONOMICS, FUTURE UNSCRIPTED: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs, (Jan. 2024), *available at* https://static1.squarespace.com/static/5ce331b47a39b9000198fffa/t/65b9314fd6198f70b0ec7402/1706635612414/Future+Unscripted+-+The+Impact+of+Generative+Artificial+Intelligence+on+Entertainment+Industry+Jobs+-+pages_compressed.pdf.

> Five… And 20 years ago, I could have just been able to live off of one of those gigs."[14]

The consulting firm Innovating With AI recently surveyed artists from a variety of media, including music, visual arts, and literature, to better understand the artists' feelings about the impact AI has had on their mediums.[15] They found that over 90% of artists view AI-generated pieces in a negative light, with no respondents saying that their view of the artwork changed positively upon learning it was AI-generated.



Source: https://innovatingwithai.com/artists-view-of-ai-generated-pieces/

The introduction of GenAI is reshaping California's creative economies. Early evidence suggests major disruption is underway, and the true impact on jobs, incomes, and long-term careers remains largely unknown. While new tools can enhance creators' workflows, the rapid influx of AI-generated content also threatens to dilute markets, suppress wages, and devalue human-made work. Further research will be necessary to fully understand the transformative effect that AI is having on California's creators.

IV. Copyright Law and Fair Use

a. Copyright law primer

The U.S. Constitution grants Congress authority to establish exclusive copyright protections through Article I, Section 8, Clause 8, which empowers lawmakers to "promote the Progress of Science and useful Arts." The stated purpose is to encourage creation and dissemination of

[14] Mozilla Foundation and the Berggruen Institute, Creativity is Collective: Hollywood's 8 Rules for AI, (Oct. 2025), *available at* https://assets.mofoprod.net/network/documents/October_2025_-_Synthesis_Report__Creativity_is_Collective__Hollywoods_8_Rules_for_AI-1.pdf.

[15] Livingston, *Over 90% of artists view AI-generated pieces in a negative light,* Innovating with AI, (Sep. 23, 2025), *available at* https://innovatingwithai.com/artists-view-of-ai-generated-pieces/.

knowledge for public benefit. Congress so used that power in establishing the Copyright Act of 1976 (the Copyright Act), which establishes copyright protections in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. (17 U.S.C. §§ 101 et seq.) It provides that all legal or equitable rights equal to those within the scope of the Copyright Act are governed exclusively by the Act, and no person is entitled to any such right or equivalent right in any such work under the common law or statutes of any State. (17 U.S.C. § 301.)

However, copyright law is "replete with escape valves: the idea-expression distinction; the general rule that facts may not receive protection; the requirement of originality; the legal standard for actionable copying; the limited duration of copyright; and . . . the defense of fair use."[16] When protected expression is copied, particularly in commercial contexts without attribution or *transformation*, questions of infringement may arise. The Copyright Act's fair use provision (17 U.S.C. § 107) identifies four factors for determining whether a given use of a copyrighted work is a "fair use":

> [T]he fair use of a copyrighted work . . . for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include —
>
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[17]

Courts have been clear that these four statutory fair use factors may not "be treated in isolation, one from another" but are "to be explored, and the results weighed together, in light of the purposes of copyright."[18]

b. Generative AI and Copyright

Relevant here and as discussed above, GenAI systems are trained on large datasets that may include copyrighted works. These systems analyze patterns from numerous sources and generate outputs that may resemble their training data in style or substance. Whether training models on copyrighted works without permission constitutes copyright infringement remains one of the

---

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

[17] *Bartz v. PBC* (2025) 787 F. Supp. 3d 1007, 1019-1020 ("*Bartz*").

[18] *Warhol*, at 550-551.

most significant unresolved issues in intellectual property law. The question involves multiple overlapping considerations about the nature of AI training, the scope of copyright protection, and the application of existing legal doctrines to new technology. This includes resolving whether training AI models on copyrighted works constitutes fair use; whether models trained on copyrighted material that produce new-form outputs constitute infringement; and how commercialization of such models affects the economic framework copyright law establishes.

AI developers often assert that training on publicly available content qualifies as transformative fair use,[19] while this position faces increasing scrutiny from courts and regulators. Some academics and AI developers argue that training is highly transformative because the purpose differs fundamentally from the original works. While a novel entertains or informs readers, using that novel to train an LLM serves the distinct purpose of teaching a model about language patterns and knowledge representation. They also argue that the training process itself does not make copyrighted works available to end users or substitute for purchasing or licensing the original works. However, others are quick to point out that LLMs sometimes reproduce portions of training data, particularly when prompted in specific ways, suggesting that protected expression has been copied and retained.

Proponents for such uses argue that models provide significant public benefits, including democratizing access to information, assisting with education and research, and enabling new forms of creativity and productivity. Some compare AI training to how humans learn by reading copyrighted works and absorbing information and patterns without infringing, provided they do not reproduce protected expression. However, many point out that unlike humans, models can process millions of works instantaneously, retain patterns with perfect consistency, and generate outputs at scale, making the human learning analogy imperfect.

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

c. Quick look at the legal landscape

While the legality of using copyrighted works for AI training remains unsettled, a number of recent cases have started to sketch out the ultimate legal landscape. A brief discussion of a few of these cases follows.

[19] Under the first factor of the statutory test, a new use is considered transformative if it "communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge." (*Authors Guild v. Google, Inc.* (2d Cir. 2015) 804 F.3d 202, 214.) Examples of relevant utility-expanding uses in prior precedents include uses have included scanning books to create a full-text searchable database and public search function (*Authors Guild, Inc. v. HathiTrust* (2d Cir. 2014) 755 F.3d 87, 97-98), copying works into a database used to detect plagiarism (*A.V. ex rel. Vanderhye v. iParadigms, LLC* (4th Cir. 2009) 562 F.3d 630, 639-640), and displaying "thumbnail" reproductions of works to provide links to websites containing the originals (*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). The transformativeness factor generally carries the most weight in determining the fairness of the use.

In *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, the Supreme Court ruled that commercial licensing of a Warhol silkscreen based on Lynn Goldsmith's photograph of Prince did not constitute fair use.[20] The majority found that despite potential aesthetic differences, the work served a similar commercial purpose. However, the dissent expressed concern that this interpretation could limit commentary, appropriation, and cultural critique.

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

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

More recently, in *Bartz v. Anthropic PBC*, the uses at issue were Anthropic's downloading for free millions of copyrighted books in digital form from pirate sites on the internet and its purchase of copyrighted books in order "to amass a central library of 'all the books in the world' to retain 'forever.'" Anthropic then used a subset of these books to train various LLMs under development. The authors sued Anthropic for copyright infringement. On summary judgment, the issue was the extent to which any of the uses of the protected works qualified as a "fair use" under Section 107 of the Copyright Act.[21] The court found the use of the books at issue to train Anthropic's GenAI model, Claude, was "exceedingly transformative and was a fair use under Section 107 of the Copyright Act." However, it found Anthropic had "no entitlement to use pirated copies for its central library. Creating a permanent, general-purpose library was not itself a fair use excusing Anthropic's piracy." The parties are currently finalizing a $1.5 billion settlement as to the use of nearly 500,000 pirated works.[22]

V. Solutions and Policy Challenges

a. Technical mitigations to promote fair use

[20] *Warhol*, 598 U.S. at 551.
[21] *Bartz*, 787 F. Supp. 3d at 1014.
[22] *Bartz v. Anthropic PBC* (2025) 2025 U.S. Dist. LEXIS 205531, *7; Kim, *Top 5 Things You Need to Know About Participating in the $1.5 Billion Bartz v. Anthropic Settlement*, Copyright Alliance (Oct. 28, 2025), *available at* https://copyrightalliance.org/participating-bartz-v-anthropic-settlement/.

Researchers have identified a number of technical approaches that can assist GenAI developers in avoiding potential copyright infringements. Henderson et al (2023) discuss five such approaches:

- *Data filtering.* Training data can be filtered to exclude likely sources of copyrighted content. In effect, developers can limit their training data to open data, excluding copyrighted or restrictively licensed materials. Additionally, developers can use filtering to remove duplicates in training data so models are less likely to "memorize" and regurgitate distinctive passages. In practice, however, this approach substantially restricts the scope of training datasets: copyrighted materials represent some of the highest quality training data available, and their exclusion potentially limits the effectiveness of the resulting models.

- *Input/Output filtering.* Some developers have added filters to their models to reject user prompts seeking to replicate an artist's style. Developers can also add filters that prevent the outputting of copyrighted content. However, such filters can often be bypassed through clever prompting.

- *Instance attribution.* Attribution scores can be assigned to training examples to assist developers in understanding how specific content contributes to model outputs. This can be used to determine the source of outputs and assign credit to creators. However, the technique is costly and may be difficult to implement at scale.

- *Differentially private training.* Differential privacy is a training technique that aims to ensure a model's behavior is not disproportionately influenced by individual copyrighted materials in the training data. This helps to prevent the model from reproducing specific content verbatim but does not address more subtle forms of infringement.

- *Learning from human feedback.* Models can be fine-tuned through human feedback, a technique that is commonly used to limit a model's harmful outputs. As humans rate outputs, the model's behavior gradually adapts to align with their preferences. This technique can be adapted to discourage models from reproducing copyrighted content.[23]

None of these techniques is a panacea for ensuring fair use. Each entails tradeoffs, such as increased costs, technical limitations, and diminished model performance. The authors of this study recommend combining multiple strategies and call for further research.

b. Identifying copyrighted materials in training data and outputs

Developers and artists can use a variety of technical strategies to identify copyrighted materials in an AI model's training data. For text, plagiarism detection tools can detect copyrighted

[23] Henderson et al, *Foundation Models and Fair Use,* arXiv (Mar. 29, 2023), pp. 20-25, *available at* https://arxiv.org/pdf/2303.15715.

passages: services and platforms like Grammarly,[24] PlagScan,[25] and Copyscape[26] can be used to scan databases to find duplicate text. Images can be checked using reverse searches – such as those offered by Google Images,[27] TinEye,[28] and Copyseeker[29] – or by generating "perceptual hashes" that match against known works. New services are emerging to address the issue of copyright violations at scale: PicDefense,[30] for example, allows the owner of a website to scan its images and compare them to billions of indexed photos to compute a "copyright risk" – the risk that the website is placing themselves in legal jeopardy by hosting copyrighted images.

Audio data can similarly be screened using fingerprinting strategies. Services like Audible Magic[31] and ACRCloud run Automatic Content Recognition (ACR) to match clips against extensive music registries. According to ACRCloud:

> [Derivative Works Detection] enables music distributors and music streaming platforms to identify copyrighted tracks in their catalogs, even when those tracks have been significantly altered through speed or pitch shifting. Using advanced audio recognition, it ensures modified works are detected, supporting copyright compliance and proper monetization while protecting original creators' rights.[32]

Videos, which are effectively long series of images paired with audio, can be checked for copyrighted materials by combining the strategies detailed above. For example, 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. Under the hood, Youtube's ContentID system pairs image and audio fingerprinting to identify likely copyright violations.[33]

Copyright owners across content modalities have also created search engines to help determine whether their works are being used to train AI. *The Atlantic*'s searchable "Books3" database allows authors to query whether their books appear in certain scraped training datasets,[34] while

---

[24] Grammarly, Plagiarism Checker, *available at* https://www.grammarly.com/plagiarism-checker.
[25] Turnitin, PlagScan, *available at* https://www.plagscan.com/en/.
[26] Copyscape, *available at* https://www.copyscape.com/.
[27] Google Lens, *available at* https://images.google.com/.
[28] TinEye, Reverse Image Search, *available at* https://tineye.com/.
[29] Copyseeker.net, AI Reverse Image Search, *available at* https://copyseeker.net/.
[30] PicDefense, *available at* https://picdefense.io/.
[31] Audible Magic, Identification, *available at* https://www.audiblemagic.com/identification/.
[32] ACRCloud, Copyright Compliance & Data Deduplication, *available at* https://www.acrcloud.com/copyright-compliance-data-deduplication/.
[33] YouTube, How Content ID Works, *available at* https://support.google.com/youtube/answer/2797370?sjid=10459553756179834900-NA.
[34] Reisner, *These 183,000 Books Are Fueling The Biggest Fight In Publishing And Tech,* The Atlantic, (Sep. 25, 2023), *available at* https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/.

"HaveIBeenTrained" offers a similar service for images included in StableDiffusion's LAION dataset.[35]

Finally, copyright owners are collaborating with researchers to develop "data poisoning" strategies. Tools such as Glaze and Nightshade add patterns to images that – while imperceptible to humans – render the images worthless for the purpose of training GenAI:

> ***Glaze*** is a system designed to protect human artists by disrupting style mimicry. At a high level, Glaze works by understanding the AI models that are training on human art, and using machine learning algorithms, computing a set of minimal changes to artworks, such that it appears unchanged to human eyes, but appears to AI models like a dramatically different art style[36] . . . ***Nightshade*** works similarly as Glaze, but instead of a defense against style mimicry, it is designed as an offense tool to distort feature representations inside generative AI image models.[37]

Copyrighted materials are more difficult to detect in AI outputs compared to training data due to AI's capacity to digest and reorganize input materials into novel configurations. Tools and techniques are being developed to overcome these hurdles, however; for text outputs, Patronus AI's CopyrightCatcher[38] allows developers to identify verbatim copyrighted passages in model outputs.

For image outputs, perceptual fingerprinting and watermarks are proving useful for detecting copyrighted materials. TraceID by Vermillio[39] generates a "neural fingerprint" for a copyrighted image, and then measures how much an AI-generated image resembles the fingerprint. Such systems essentially calculate high-level overlap in visual features between an orginal work and an output. Researchers have proposed using these types of perceptual hashes to perform comparisons that are robust to common edits such a cropping, resizing, or changing the brightness or contrast of an image.[40] For audio, Pex offers phonetic lyric matching – designed to catch unauthorized lyric use – as well as voice-matching to spot particular artists' voices in outputs:

> Pex's phonetic fingerprinting and matching enables the identification of two different audio files that contain the same underlying words, regardless of whether they are sung or spoken. Our technology analyzes the underlying phonetic content of the audio recording itself (both individual words and combined phrases) and

---

[35] Kemper, *A new website makes AI training images searchable,* the decoder, (Oct 17, 2022), *available at* https://the-decoder.com/a-new-website-makes-ai-training-images-searchable/.
[36] The Glaze Project, Glaze, *available at* https://glaze.cs.uchicago.edu/.
[37] The Glaze Project, Nightshade, *available at* https://nightshade.cs.uchicago.edu/.
[38] Patronus AI, Introducing CopyrightCatcher, the first Copyright Detection API for LLMs, (Mar. 6, 2024), *available at* https://www.patronus.ai/blog/introducing-copyright-catcher.
[39] Milmo, *The platform exposing exactly how much copyrighted art is used by AI tools,* The Guardian, (Oct. 18, 2025), *available at* https://www.theguardian.com/technology/2025/oct/18/the-platform-exposing-exactly-how-much-copyrighted-art-is-used-by-ai-tools.
[40] TrufoAI, *A New Era of Perceptual Hashing,* Medium, (Sep. 28, 2025), *available at* https://medium.com/trufo/a-new-era-of-perceptual-hashing-fce2b7e04591.

> creates a unique representation of the phonetics, independent of the unique vocal style and cadence being used by the singer or speaker.[41]

In summary, a variety of tools have been developed to help identify copyrighted materials in AI training datasets and output. As GenAI continues to proliferate, this ecosystem of detectors, protocols, and rights registries will continue to expand.

c. How other countries are grappling with AI and copyright

In a report on copyright and GenAI training, the US Copyright Office included the following discussion of how other countries have addressed use of copyrighted works to train AI models:

> In the EU, the 2019 Directive on Copyright in the Digital Single Market (DSM Directive) directs member states to provide exceptions for "reproductions and extractions" of copyrighted material for use in TDM [text and data mining], in certain circumstances. Article 3 of the DSM Directive applies only to TDM activities by "research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access." Article 4 is broader and applies to TDM activities by any actor for any purpose, but conditions the availability of the exception on lawful access and respecting opt-outs by copyright owners.
>
> In 2024, the EU adopted the Artificial Intelligence Act ("EU AI Act"), which references the DSM Directive's TDM exceptions in the context of generative AI. Recital 105 acknowledges that TDM techniques "may be used extensively in [the context of training AI models] for the retrieval and analysis of such content, which may be protected by copyright and related rights." Article 53 obligates AI model providers to establish policies for complying with Union law and to identify and comply with copyright owner opt-outs under the DSM Directive's Article 4 TDM exception.
>
> There continues to be controversy, however, over how the TDM exceptions apply to uses involving generative AI and whether and how the opt-out provision will work. Discussions continue at both the EU level and in member states, and so far there is little case law on point. At this stage, it remains to be seen how that opt-out provision will be implemented by individual EU member states.
>
> In other jurisdictions as well, various limitations or conditions have been included in TDM exceptions. Singapore's version requires lawful access to the work and limits the use of copies to the purpose of computational data analysis. Copies may only be supplied to others for the purposes of verifying results or collaborative research.

[41] Galka, Identify more compositions and AI-generated voices with Pex phonetic matching, (Nov. 27, 2023), *available at* https://pex.com/blog/identify-more-compositions-and-ai-generated-voices-with-pex-phonetic-matching/.

Japan's TDM exception allows the use of a copyrighted work for AI development or other forms of data analysis as long as the use is not to "personally enjoy…the thoughts or sentiments expressed in that work." The exception does not apply if "the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation." In its 2024 AI guidelines, Japan's Copyright Office explained that "enjoyment" refers to "the act of obtaining the benefit of having the viewer's intellectual and emotional needs satisfied through using the copyrighted work," citing examples such as reading literary works, appreciating musical works, and executing works of computer programming. Generating material similar to the original works can be "for enjoyment," and if a user's purpose is even partly for enjoyment, the exception does not apply. Similarly, "reproducing a copyrighted database work for the purposes of data analysis, such as AI training for which licenses for data analysis are available in the marketplace," is not covered.

UK law contains a narrower exception, dating back to 1988, that permits copying to "carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose," but only if the copier has lawful access to the work. As part of its recent consultation on Copyright and Artificial Intelligence, the government has inquired into the application of this exception to AI and sought comments on introducing a TDM exception subject to copyright owner opt-outs, similar to the approach in the EU. This proposal has proved quite controversial, with commenters warning that it would impose burdensome transaction costs for both copyright owners and AI developers.

Other countries have approached the legal status of AI training through the lens of fair use. In Israel, the copyright law includes a provision closely modeled on section 107 of the U.S. Copyright Act. In December 2022, the Ministry of Justice released an Opinion on the uses of copyrighted materials for machine learning, concluded that the use of copyrighted materials in machine learning datasets and training process is, in most but not all cases, fair use. It cautioned, however, that the Opinion "does not apply to [machine learning]-based products, but only to the learning process itself. The infringing status of the product will be examined adhoc based on extant copyright rules and standards, and this Opinion does not grant products an a-priori safe harbor."

In Korea, the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission in 2023 released *A Guide on Generative AI and Copyright*. The guide recognizes that there is "an ongoing debate within academia on the applicability of the fair use rule" and observed that until "several related court precedents accumulate," the "applicability of the fair use defense will remain unclear," leaving open the possibility that "using a work for AI training without permission from the copyright holder" may constitute infringement.

Approaches to generative AI and copyright matters in the People's Republic of China are developing, and it is not yet clear how the use of copyrighted works in training will be treated. The Copyright Act does not have an express exception for text and data mining activities or AI training. Article 24 of the Act contains a list of enumerated exceptions, including a new open-ended exception covering "other circumstances as provided in laws and administrative regulations." With respect to litigation, one recent case held an AI platform provider contributorily liable for infringements occurring when users uploaded protected content into models available via the platform, which generated infringing copies. While there have been other cases involving infringing output, it appears that courts have yet to consider a copyright infringement claim against a foundation model developer based on the use of copyright protected works to train a foundation model. Meanwhile, press reporting on the annual work report from the Supreme People's Court indicates that the issue of intellectual property and AI is an area of ongoing attention. China has also issued at least two administrative measures providing guidance on generative AI services, including compliance requirements for training data. Avenues for supporting and developing the AI sector were topics receiving significant press coverage in relation to the March 2025 National People's Congress.

Finally, a few countries are considering statutory approaches to compensation. In Brazil, a pending bill would require AI companies to compensate rightsholders for the use of their works in training. The draft directs the parties to discuss compensation in a manner that allows rightsholders to negotiate effectively either directly or collectively, calculate compensation that reasonably and proportionally considers the AI agent's size and the potential competition impacts; and preserves freedom of agreement. In 2024, Spain opened public commentary on a Draft Royal Decree which would establish an extended collective licensing mechanism for the mass exploitation of protected works in the development of AI models, although the proposal was subsequently withdrawn.[42]

[42] Copyright and Artificial Intelligence, Part 3: Generative AI Training, U.S. Copyright Office (May 2025), pp. 76-82, footnotes omitted, *available at* https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf.