VOLUME 31 SUMMER 2026

I L L I N O I S B U S I N E S S L A W J O U R N A L

NAVIGATING THE SECTION 1202(B) DIVIDE: THE NECESSITY OF LENIENT PLEADING STANDARDS IN AI COPYRIGHT DISPUTES

❖ Note ❖

*Chen-Huan Wang**

I. INTRODUCTION

In today's digital era, original creative works are seamlessly integrated into the online ecosystem, where creators routinely embed signatures, titles, and copyright notices directly onto their digital media. Consider an artist who uploads her digital illustration online, with her signature and copyright notice displayed alongside the image. Months later, an Artificial Intelligence company collects millions of images to train a generative AI model. During that data collection process, AI developers are alleged to use tools that separate the underlying work from accompanying copyright information, such as copyright notices, author names, and other identifying information.[1] Even if the resulting AI model may not reproduce that exact image in its outputs, it has nevertheless been trained on a version of the work stripped of the information identifying its creator.[2] Consequently, this raises the question of whether copyright law should recognize this as a statutory violation.

This Note dives into the "black box" of AI training, specifically focusing on intentional removal or alteration of Copyright Management Information (CMI) during the AI training process. When AI developers systematically erase these critical "identifiers" from a work, they fundamentally strip authors of their attribution rights, thus threatening to undermine the core incentives for future societal innovation.[3]

* J.D. Candidate, Class of 2027, University of Illinois College of Law

[1] *See N.Y. Times Co. v. Microsoft Corp.*, 777 F. Supp. 3d 283, 314–15 (S.D.N.Y. 2025).
[2] *Id*.
[3] *Harper & Row, Publrs. v. Nation Enters.*, 471 U.S. 539, 546, 105 S. Ct. 2218, 2223 (1985).

By exploring Section 1202(b) of the Digital Millennium Copyright Act (DMCA) and relevant case law, this Note navigated the emerging judicial split over how plaintiffs can plausibly allege the statutory scienter requirements, given the inherently opaque nature of AI development. As the scraping and training phases operate within a proprietary "black box," courts have diverged on the evidentiary burden required to survive a motion to dismiss. While some jurisdictions enforce a stringent standard that penalizes the plaintiff's lack of insider technical knowledge, others apply a more lenient standard that acknowledges these informational barriers. At its core, this conflict centers around the creator's attribution rights. When developers utilize tools to strip CMI, they dismantle the legal framework that has protected authors for decades. Consequently, this Note advocates for a more lenient pleading standard regarding the plaintiff's burden of proof. To support this thesis, Part II examines the foundational significance of CMI and pleading challenges created by the AI training "black box". Part III analyzes the current jurisprudential divergence, contrasting the output-driven, restrictive approach with the process-oriented, lenient standard. Part IV justifies the necessity of this lenient framework by drawing an analogy to the doctrine of res ipsa loquitur to address information asymmetry and exploring the legislative impact of California's Generative AI Training Data Transparency Act. Ultimately, while this more lenient standard may increase litigation, it is necessary to ensure a transparent AI ecosystem that fosters development and furthers innovation in the long run.

II. BACKGROUND

A. The Significance of CMI to Authors

Congress enacted the DMCA in 1998 with the legislative intent "to strengthen copyright protection in the digital age," and to combat copyright piracy.[4] At the core of this legislation is Section 1202(b), which explicitly prohibits intentional removal or alteration of copyright management information.[5] As defined in Section 1202(c), Copyright Management Information (CMI) is the information conveyed in connection with copies or phonorecords of a work or performance, including in digital form.[6] It includes the name of, and other identifying information about, the author of the work[7], and acts as an "identifier."

Functioning as a critical digital "identifier", CMI is significant to authors as it protects attribution rights, ensures that the author receives proper credit when

[4] *N.Y. Times Co. v. Microsoft Corp.*, 777 F. Supp. 3d 283, 312 (S.D.N.Y. 2025).
[5] 17 U.S.C. § 1202(b) (2018).
[6] 17 U.S.C. § 1202(c) (2018).
[7] 17 U.S.C. § 1202(c)(2) (2018).

referenced, and allows the author to monitor how their works are utilized.[8] The unauthorized removal of CMI effectively "anonymizes" an author's work, stripping away their attribution right. Consequently, the absence of CMI significantly hinders an author's capability to prove copyright infringement, thereby significantly undermining the statutory purpose of this Act which is designed to protect authors.

Generative AI models, including Large Language Models (LLMs), rely on the ingestion of massive volumes of data to train their systems, much of which is derived from datasets "scraped" from the internet.[9] However, AI developers typically do not disclose the specific contents of their training datasets or their scraping methodologies.[10] This lack of transparency is known as the "black box problem" of AI training.[11] As the system's internal decision-making process is obscured, it becomes nearly impossible to trace the AI's thought processes, determine the rationale behind its generated decisions, or understand precisely how and why specific data was utilized.[12] Operating under the "black box", developers are increasingly accused of systematically stripping away CMI during their training process and thus triggering the Section 1202(b) violation.

B. Section 1202(b)

Section 1202(b)(1) states "No person shall, without the authority of the copyright owner or the law— (1)intentionally remove or alter any copyright management information, knowing, or, with respect to civil remedies under section 1203, having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement of any right under this title."[13]

To establish liability under this framework, Section 1202(b)(1) contains a "double-scienter" requirement. Specifically, a plaintiff must demonstrate that a defendant not only (1) "intentionally" remove CMI under Section 1202(b)(1), but also (2) know or have reason to know that its conduct would induce, enable, facilitate, or conceal infringement.[14]

Since the AI training process operates as a "black box", proving this double-scienter requirement presents a legal challenge. In determining whether AI developers have violated Section 1202(b) during this obscured scraping phase,

---

[8] Megan Keenan, *Copyright Management Information (CMI) as a Tool to Protect Indigenous Cultural Works*, 12 N.Y.U. J. Intell. Prop. & Ent. L. 413, 413 (2023).

[9] Cong. Rsch. Serv., R47569, *Generative Artificial Intelligence and Data Privacy: A Primer* 3 (2023).

[10] Lou Blouin, *AI's Mysterious "Black Box" Problem, Explained*, U. Mich.-Dearborn News (Mar. 6, 2023), https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained.

[11] *Id.*

[12] *Id.*

[13] 17 U.S.C. § 1202(b)(1) (2018).

[14] *Id.*

courts have diverged on factors that establish this violation. Some jurisdictions adopt a relatively stringent approach that increases the plaintiff's burden of proof during the pleading stage, while others apply a more lenient standard that facilitates the plaintiff's ability to state a plausible claim.

III. ANALYSIS

The judicial application of Section 1202(b) to generative AI reflects divergence over the burden of proof necessary to satisfy the "double-scienter" requirement of Section 1202(b) at the pleading stage. At its core, the courts split on whether a violation occurs primarily during the "input" phase, when CMI is stripped to build AI training datasets, or only during the "output" phase, when the generated AI content reproduces the original work without its identifiers.[15] Notably, this legal divergence mirrors a geographical split. Some courts in the Northern District of California enforce a restrictive, output-driven standard that requires proof of an "identical output" and identification of the exact CMI scraped.[16] Conversely, other courts in the Southern District of New York adopt a more lenient, process-oriented standard, allowing claims to proceed if plaintiffs simply show that developers intentionally utilized algorithms to discard CMI during the training phase.[17]

The divergence creates a hurdle for plaintiffs, as the standard adopted by a court often dictates whether a creator can move past a motion to dismiss or is barred by an information gap inherent to the proprietary AI models. Ultimately, these judicial thresholds signal the level of protection a creator can expect in the digital age, which may profoundly influence the long-term incentive to innovate and produce novel original works.

A. Restrictive Approach Focusing on Producing the Identical Output

Some courts enforce a highly restrictive, output-driven standard that demands plaintiffs identify the exact CMI scraped and prove the AI generated an "identical output" to the original work.

---

[15] *Compare Doe v. GitHub, Inc*., No. 22-cv-06823-JST, 2024 U.S. Dist. LEXIS 11068, at *24–25 (N.D. Cal. Jan. 3, 2024), *and Andersen v. Stability AI Ltd*., 700 F. Supp. 3d 853, 872 (N.D. Cal. 2023), *with* N.*Y. Times Co. v. Microsoft Corp*., 777 F. Supp. 3d 283, 314–15 (S.D.N.Y. 2025), *and In re OpenAI, Inc. Copyright Infringement Litig*., No. 25-md-3143 (SHS) (OTW), 2025 U.S. Dist. LEXIS 259061, at *72–73 (S.D.N.Y. Dec. 15, 2025).

[16] See *Doe v. GitHub, Inc.*, No. 22-cv-06823-JST, 2024 U.S. Dist. LEXIS 11068, at *23–25 (N.D. Cal. Jan. 3, 2024)*; Andersen v. Stability AI Ltd., 700 F. Supp. 3d 853, 872 (N.D. Cal. 2023).*

[17] See *N.Y. Times Co. v. Microsoft Corp.*, 777 F. Supp. 3d 283, 314–15 (S.D.N.Y. 2025); *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 25-md-3143 (SHS) (OTW), 2025 U.S. Dist. LEXIS 259061, at *53–55, 72–73 (S.D.N.Y. Dec. 15, 2025).*

In *Doe v. Github*, software developers argued that the development and operation of AI-based coding tools, Copilot and Codex, by GitHub, Microsoft, and OpenAI were trained on billions of lines of licensed code from public repositories[18] but were not programmed to respect open-source license requirements, such as attribution and copyright notices.[19] Plaintiffs argued that the Defendants intentionally designed their programs to remove or alter CMI during the training and output process, thereby violating Section 1202(b).[20] The court determined that Section 1202(b) claims are only viable when CMI is removed from an "identical copy" of a copyrighted work.[21] Since the plaintiff's complaint admitted that Copilot output is often a "modification" or a "near-identical copy" rather than an identical reproduction, the court dismissed the Section 1202(b) claims, concluding that mere similarity is insufficient to sustain a DMCA violation under this provision.[22] Here, the court applies a restrictive standard, holding that a violation only exists if CMI is removed from an "identical copy," and mere modifications are not enough.

Similarly, in *Andersen v. Stability AI Ltd.*, several artists (plaintiffs) alleged that generative AI companies, including Stability AI, DeviantArt, and Midjourney (defendants), trained their AI models on their copyrighted works after CMI was illegally removed.[23] The plaintiffs contended that the removal of identifying information, such as artists' signatures and creator names, during the scraping and training process violated Section 1202(b) of the DMCA.[24] However, the court found these allegations to be "wholly conclusory" and dismissed the claim.[25] The court's reasoning focused on the plaintiff's failure to identify the specific CMI included in the particular works that were actually scraped, as well as their failure to provide plausible facts showing which specific defendant performed the stripping of CMI and at what point in the training process it occurred.[26] Here, the court similarly applied a restrictive standard requiring plaintiffs to identify the exact type of CMI included in specific works that were allegedly scraped.[27]

B. *Lenient Approach Focusing on Systemic Input*

[18] *Doe v. GitHub, Inc.*, 672 F. Supp. 3d 837, 846 (N.D. Cal. 2023).
[19] *Id.* at 847.
[20] *Id.* at 857.
[21] *Doe v. GitHub, Inc.*, No. 22-cv-06823-JST, 2024 U.S. Dist. LEXIS 11068, at *24–25 (N.D. Cal. Jan. 3, 2024).
[22] *Id.*
[23] *Andersen v. Stability AI Ltd.*, 700 F. Supp. 3d 853, 860 (N.D. Cal. 2023).
[24] *Id.* at 872.
[25] *Id.*
[26] *Id.*
[27] *See Id.* at 871.

Conversely, other courts have begun to recognize the "black box" nature of AI development, embracing a more lenient, process-oriented approach. These courts find claims plausible if plaintiffs can demonstrate that AI developers intentionally utilized algorithms to extract text while purposefully discarding CMI to build their training datasets or conceal infringement.

In *N.Y. Times Co. v. Microsoft Corp.*, the various news organizations (plaintiffs) alleged that Microsoft and OpenAI (defendants) violated 17 U.S.C. Section 1202(b) by intentionally removing or altering CMI during the training of their LLMs.[28] Specifically, the plaintiffs contend that the defendants utilized training datasets that stripped author and title information from their original works.[29] While the Court dismissed the claims against Microsoft and the New York Times action against OpenAI, it allowed the Daily News and Center for Investigative Reporting (CIR) claims to proceed against OpenAI.[30] The court's reasoning centered on the specificity of pleadings; whereas the Times failed to provide detail beyond the fact that CMI did not appear in AI outputs, the Daily News and CIR plausibly alleged that OpenAI intentionally used specific algorithms, Dragnet and Newspaper, designed to "separate and extract" article text while purposefully discarding CMI like footers and copyright notices to ensure dataset consistency.[31] The court found that the Daily News and CIR satisfied the "double-scienter" requirement at the pleading stage by showing OpenAI likely knew such removal would facilitate or conceal its own infringement.[32]

Similarly, in *Davis v. OpenAI, Ziff Davis*, a major digital media publisher, alleged that OpenAI systematically used its copyrighted articles without authorization to train large language models (LLMs) and intentionally removed CMI during the scraping and training process.[33] The court denied OpenAI's motion to dismiss[34], relying on its prior reasoning in the New York Times case. The court found that Ziff Davis successfully stated a claim by alleging that OpenAI reproduced and created identical or derivative versions of its works while intentionally stripping CMI to build training datasets.[35] At the pleading stage, the court emphasized that the plaintiff only needed to establish a "plausible" allegation[36], noting that Ziff Davis provided documented examples of ChatGPT generating verbatim outputs of its works without CMI is enough to allege such

---

[28] *N.Y. Times Co. v. Microsoft Corp.*, 777 F. Supp. 3d 283, 310 (S.D.N.Y. 2025).

[29] *Id.* at 315.

[30] *Id.* at 316.

[31] *Id.* at 314–15.

[32] *Id.* at 315.

[33] *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 25-md-3143 (SHS) (OTW), 2025 U.S. Dist. LEXIS 259061, at *53–54 (S.D.N.Y. Dec. 15, 2025).

[34] *Id.* at *76.

[35] *Id.* at *72–73.

[36] *Id.* at *55.

violations.[37] Here, the court applied a more lenient standard that only requires the plaintiff to show a "plausible capacity" for the AI to generate verbatim copies without CMI.

IV. RECOMMENDATION

A. The Necessity of a Lenient Pleading Standard

As evidenced by the four cases discussed above, courts have applied divergent standards regarding what constitutes a Section 1202(b) violation of the DMCA. In *Doe v. GitHub and Andersen v. Stability AI Ltd.*, the courts adopted a restrictive approach, requiring plaintiffs to demonstrate that the AI produced an "identical" output to the copyrighted work.[38] In contrast, in *N.Y. Times Co. v. Microsoft Corp*. and *Davis v. OpenAI*, the courts found allegations plausible based on the "process", specifically, by proving that the Generative AI companies stripped CMI from the copyrighted work during the AI training phase, without strictly requiring an "identical" output.[39]

This note contends that the more lenient standard applied in N.Y. Times and Davis is more appropriate. If the restrictive view requiring identical output is applied, the burden of proof for plaintiffs to meet the plausibility standard is heightened. Such a heightened standard makes it increasingly difficult for authors and creators to protect their copyrights, which might ultimately hinder their incentive to innovate and create. Moreover, under a restrictive standard, users of generative AI might unknowingly infringe upon copyrighted works without the opportunity to properly credit the original authors.

Furthermore, these restrictive standards are inappropriate given the inherent information disparity in AI litigation.[40] In the legal doctrine "Res Ipsa Loquitur," particularly within civil procedure, courts sometimes apply a burden-shifting framework when a profound information asymmetry exists between parties. A classic example can be found in medical malpractice litigation, where a patient who is often unconscious and lacks specialized knowledge cannot locate the exact moment of negligence, while the hospital and medical staff possess exclusive control over the records essential to proving negligence in the case[41]. Analogous to this doctrine, many creators are unfamiliar and lack the specialized engineering knowledge of how AI companies train their models. Whereas, AI developers

[37] *Id.* at *73.

[38] *Doe*, 2024 U.S. Dist. LEXIS 11068, at *23.

[39] *N.Y. Times Co.*, 777 F. Supp. 3d at 315.

[40] *See Swierkiewicz v. Sorema N.A.*, 534 U.S. 506, 512 (2002).

[41] *See* Restatement (Third) of Torts: Medical Malpractice § 7 (Am. L. Inst., Tentative Draft No. 2, 2024).

possess exclusive control over the proprietary information of their datasets and scraping methodologies that are largely inaccessible to plaintiffs prior to discovery.[42] While the court may not explicitly shift the burden of proof to the defendants, it should at least adopt a more lenient pleading standard to ensure that plaintiffs have a procedurally fair opportunity to state their case and move beyond the motion to dismiss stage.

Furthermore, adopting a relatively lenient standard provides critical protection for the users of generative AI tools. When CMI is preserved throughout the training and output phases, users gain the ability to filter and verify information based on its original source. This human filtering process is essential in an era of AI-generated misinformation, but by knowing the source of a particular data point, users can better assess the accuracy and trustworthiness of the output instead of an anonymized result that lacks the context necessary for professional or academic validation. This empowers the user to distinguish between credible, high-quality data and unreliable content, ultimately delivering a more transparent and reliable digital ecosystem.

B. The California Generative AI Act

The recent passage of California's new AI Law- Generative AI Training Data Transparency Act (AB 2013), intersects directly with the restrictive pleading standards currently applied in the federal courts within California. This bill, effective January 1, 2026, requires AI developers to publicly disclose high-level summaries of their training data used in the development of generative AI systems or services.[43] Under Section 3111 of this bill, the high-level summary should include the number and description of datasets, whether the datasets include copyrighted or personal information, and whether there were any cleaning, processing, or other modifications to the datasets by the developer.[44] By forcing developers to reveal what datasets were used, it could be easier for authors to identify the specific works scraped, potentially helping them survive a motion to dismiss under the restrictive standard that demands plaintiffs identify the exact CMI removed.

However, this legislative push for transparency doesn't preclude the adoption of a more lenient standard for two reasons. First, this act only requires high-level "summaries" of training data, meaning it still fails to provide the plaintiffs with which "specific" CMI was stripped to meet the courts' standard. Second, the plaintiffs are still required to prove that the AI generated an "identical output" of the original work. Thus, even with this act, the courts can still dismiss the claim if

[42] *See Andersen,* 700 F. Supp. 3d at 872–73.
[43] Assemb. B. 2013, 2023–2024 Reg. Sess. (Cal. 2024) (Legislative Counsel's Digest).
[44] *Id.*

the AI output is a "modification" rather than an identical reproduction. Therefore, to truly protect authors, courts should still adopt the lenient, process-oriented standard that recognizes violations during the systemic input phase.

Some might argue that a more lenient approach could hinder or deter AI innovation by imposing burdensome compliance requirements. However, this proposal simply asks generative AI companies to refrain from actively stripping CMI during the scraping phase, and merely train their AI models in the datasets as given. In the long term, society is better served by a legal ecosystem that respectfully protects authors. By ensuring the integrity of attribution, the law fosters more robust innovation by maintaining a fair and sustainable marketplace for human and artificial creativity.

V. CONCLUSION

As the legal landscape surrounding AI continues to shift, the judicial interpretation of Section 1202(b) will serve as a gatekeeper for creators' rights. Without a clear and definite standard as to what satisfies the "double-scienter" requirement, the current geographic jurisdictional split between the Northern District of California and the Southern District of New York might incentivize strategic "forum shopping," where AI developers gravitate toward west coast jurisdictions that uphold a restrictive "identical output" standard to shield their proprietary training processes from scrutiny.

The law must ultimately recognize that AI training is built upon an immense foundation of human creativity. Allowing the intentional stripping of CMI can effectively sever the essential link between authors and their work, which undermines the very incentive that drives cultural innovation. By adopting a more lenient standard that protects the integrity of attribution during the training phase, the judiciary can foster a fair and sustainable marketplace that ensures AI augments, rather than erases the human creators who make its existence possible.