AI watermarks erase human authors twice, while claiming to protect authenticity

Published on September 1, 2026

Models absorb human creativity without consent, then label human-directed work as machine-made, creating a troubling double erasure of authorship.

A system built to flag synthetic content ends up flattening every degree of human involvement into one undifferentiated "machine-made" category. Wikimedia Commons, Creative Commons Attribution-Share Alike 4.0 International, CC-BY-SA-4.0.

Authors Elizabeth Lyn
O.P. Jindal Global University, Sonipat

Editors Chandan Nandy
Commissioning Editor, 360info
Namita Kohli
Commissioning Editor, 360info

Every time someone asks an AI tool to draft an email, polish an essay or write a report, an invisible marker gets attached to the result, classifying it as "AI-generated." It discounts the human who wrote the prompt, steered the revisions and approved the final version.

It also sits on top of a deeper problem: these models learned to write by ingesting decades of human writing, much of it without consent or payment. Humans are erased twice: once when their past work trains the model without credit, and again when their present work gets tagged as not quite theirs. That double erasure is the real story.

Text watermarking, at least, was not the industry's idea; regulators forced it. The European Union's AI Act requires generative AI providers to mark their output in machine-readable form. The goal is reasonable: as AI deepfakes and synthetic media spread, regulators wanted platforms and the public to be able to tell machine-made content from human-made content.

But the tools built to serve that goal do something cruder than advertised. Text watermarks such as Google's SynthID work by subtly biasing which words a model chooses during generation, embedding a statistical fingerprint in the phrasing. Anthropic's Claude models now do the same.

This fingerprint cannot distinguish a one-line prompt from months of human-directed drafting. Ask a model to lightly proofread an essay you wrote entirely yourself, and the finished text can still carry the mark, because detection only shows the model touched the text, not how much of the thinking was yours. A system built to flag synthetic content ends up flattening every degree of human involvement into one undifferentiated "machine-made" category.

Money does not buy an exemption either. No Claude plan, free or paid, offers a way to switch the watermark off. If payment and provable authorship both fail to move the needle, the system was never built around crediting the human in the loop.

And even once detection tools exist, they will not answer the question anyone actually cares about. Anthropic itself says a detected mark is not conclusive evidence Claude produced the content, and the absence of a mark cannot guarantee AI was not involved either. The most a detector can say is that a Claude model touched a piece of text. It cannot say whether that touch was a full first draft or a single comma fixed in someone else's finished essay, or weigh whose ideas shaped the argument.

Machine-readable mark

There is another catch. Because the mark is machine-readable rather than visible, it was never meant to be what a reader sees directly. Under the EU's rules, a platform cannot simply treat an AI company's hidden mark as its own public warning; a separate visible label may be required. The mark meant to help the public spot synthetic content cannot reach the public without infrastructure that is still developing.

The signal is easy to erase. Anthropic itself acknowledges that a thorough rewrite, one where every word is replaced, will strip the fingerprint out, while light editing probably will not. The technology ends up catching the people least likely to cause harm while doing little to stop the ones who are.

The criticism is less about bad faith than about a transparency measure whose honest disclaimers sit awkwardly next to a policy with no exceptions and no opt-out.

Now turn to the other end of the process, where the erasure runs the opposite direction. Large language models did not learn to write from nothing. They were trained on an almost unimaginable volume of human journalism, books, code and art, much of it absorbed without consent or payment.

Copyright clash

The New York Times is suing OpenAI and Microsoft, alleging their models can reproduce Times' articles nearly verbatim; that case remains pending and unresolved, so the allegation should be read as a claim, not a finding. Anthropic, separately, settled a case over allegations it downloaded millions of pirated books to train Claude.

A federal judge approved a $1.5 billion settlement in July 2026. But the legal distinction matters: the underlying ruling found that using legally acquired books to train an AI model was fair use. What exposed Anthropic to liability was downloading and permanently storing pirated copies, not the act of training on them.

In India, news agency ANI sued OpenAI in the Delhi High Court over using its content to train ChatGPT. The court denied ANI's request for an interim injunction in July 2026, ruling on a prima facie basis that the training use fell within India's fair-dealing exception.

The broader point is that the law remains unsettled. In the US, the New York Times' case against OpenAI and Microsoft remains unresolved, while other cases have shown models capable of reproducing substantial chunks of original text. Yet no court, anywhere, has issued a final, binding ruling that AI training itself is copyright infringement. The law is being written now, not settled in either direction.

Seen together, the pattern is still worth naming, even with those caveats attached. At the input stage, companies leaned on fair use to absorb the uncompensated work of millions of writers and artists, work that plausibly constitutes real creative substance behind what these tools produce, even where courts have so far permitted the practice.

At the output stage, the same companies attach a classification that discounts the human who directs and edits the result, applied regardless of payment tier. Credit is denied coming and going.

None of this argues against disclosure itself. Readers deserve to know when they might be looking at synthetic media. But a watermark that cannot distinguish a lazy prompt from months of human judgment, that a thorough rewrite can strip out entirely, and that mostly cannot reach the public it was built to protect, is not really tracking authorship.

It is tracking something far shallower and calling it by a bigger name. Fixing that means a more honest, graded account of human involvement rather than a single binary tag, paired with equally honest terms for the human work that trained these systems in the first place, including whether the people who supplied that labour were ever compensated for it.

Until regulators address both ends of this problem together, the "AI-generated" tag will keep doing far more classifying than the underlying facts, or the underlying law, can currently justify.

Elizabeth Lyn *is a Lecturer at the Jindal School of Government and Public Policy, O.P. Jindal Global University, Sonipat, Haryana.*