PRODUCTIVITY · 15 min read · August 18, 2026

Remove Claude's Watermark for Free

Claude now marks everything it writes. Here's the free way to remove it, how the mark actually works, what it can and can't prove, what the law requires of you, and what students need to know about false accusations.

THE FREE METHOD, RIGHT UP FRONT

Rewrite the text in your own words. That breaks the mark. Claude's watermark lives in the specific words it chose, so paraphrasing it, translating it, or blending it with your own writing all destroy the pattern. No install, no signup, no cost. That is the whole method, and it's exactly what the GitHub tool automates by having a second AI do the rewording for you.

The tool is here: github.com/guillaumemeyer/watermarks-remover. It's free and open-source. It earns its keep on volume, and on two jobs hand-editing can't do: stripping the invisible Unicode characters AI leaves in your text, and clearing the hidden data attached to AI-generated images. Both are real, both are fast, and the section on that tool below covers what its makers admit it can't promise.

And you have more time than the panic suggests. The detection tools aren't public yet, not to your client, not to your school, not to anyone. So you can handle this properly instead of rushing to install something tonight.

In August 2026, Claude started marking everything it writes. You can’t see the mark. You can’t turn it off. And it’s on every Claude product, everywhere in the world, not just in Europe.

That set off a lot of panic online, and most of the coverage swung to one extreme or the other. Either you’re about to get caught, or you need to go install something immediately. The real picture is calmer than both, and knowing how the mark is built is

what lets you deal with it in about a minute.

The internet's version of this story, and a good example of the problem. That confident "98.7% detected" readout is exactly the kind of claim being made right now. As section 4 covers, no such detector is available to anyone yet.

Here's the real thing: what changed, how the mark is built, what it genuinely can and can't detect, the one part of this law that applies to you and not to Anthropic, and what students in particular need to know, because the risk there isn't the one most people are worried about.

1 What actually changed

The European Union passed a law called the AI Act. One piece of it, Article 50, covers transparency, meaning it forces AI companies to be honest about what's AI-made. That piece became enforceable on August 2, 2026.

The rule for AI companies is short: if your system generates text, images, audio, or video, the output has to be marked in a machine-readable format so it can be detected as artificially generated. Not visible to a human necessarily, but readable by a machine that knows what to look for.

The penalty for ignoring it is real money. Up to 15 million euros, or 3% of a company's total worldwide annual revenue, whichever number is bigger.

Anthropic signed on. Starting August 2, 2026, every Claude model released from that date forward marks its output. Older Claude models have until December 2, 2026 to catch up.

IN PRACTICE

The law is European. The marking is not. Anthropic applies it worldwide, across every place a supported model runs: the Claude apps, the API, Claude Code, and Claude running on Amazon, Google, and Microsoft's cloud platforms. It was simpler to mark everything than to build one version for Europe and one for everyone else. So it applies to you whether you live in the EU or not.

2 How the text mark actually works

Understand this part and everything else in the guide follows from it.

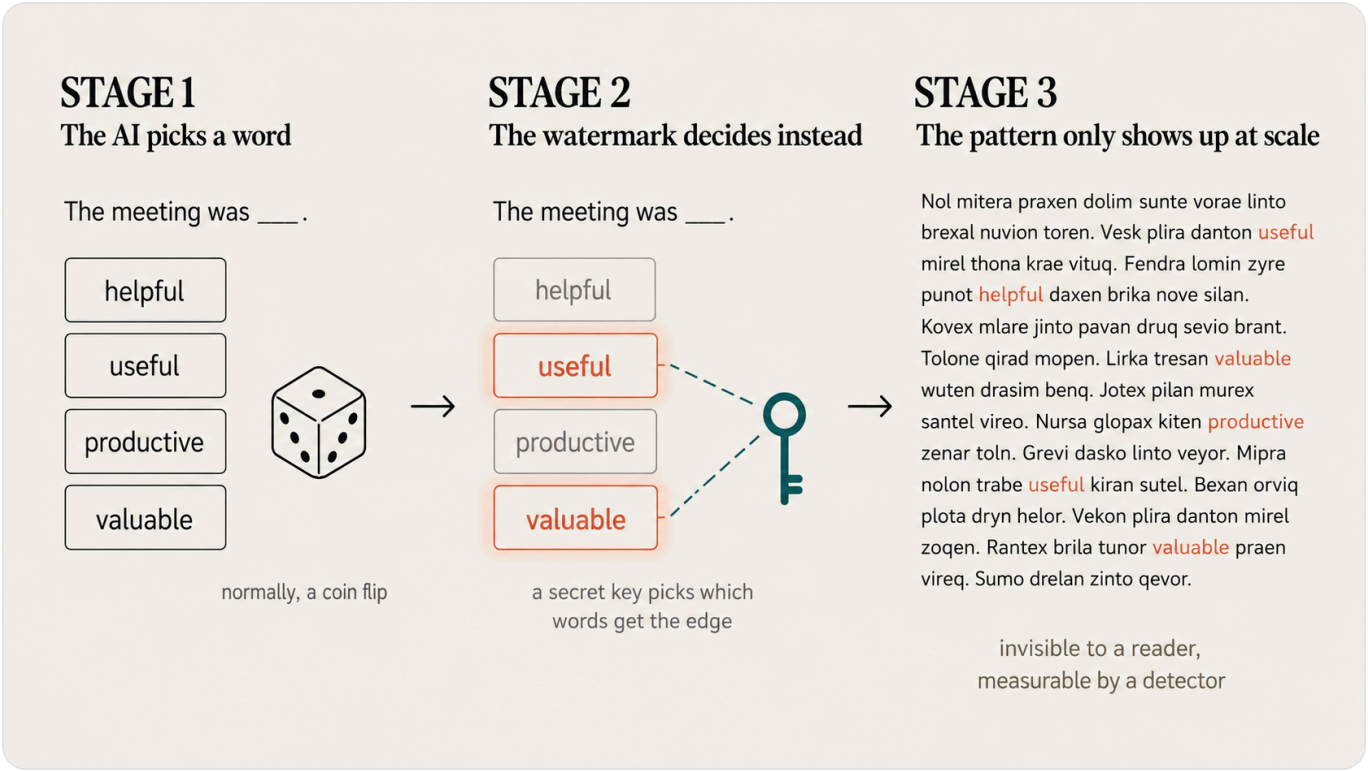
The mark isn't hidden characters. It isn't invisible spaces or weird punctuation you could find and delete. It's built into the word choices themselves.

Here's the mechanism. Every time an AI writes, it's picking the next word from a list of good options. Usually several words would work fine. "The meeting was helpful" and "the meeting was useful" are both natural, and the model is flipping a coin between them.

The watermark takes over that coin flip. A secret key decides, in a pattern only that key can recognize, which of the good options gets picked. The writing still sounds normal, because every word chosen was already a reasonable word. But across a few hundred words, the pattern of those choices lines up in a way that random chance would almost never produce.

A detector holding the same key can measure that pattern and say “this looks like our signature.” A human reading it sees nothing at all.

Anthropic didn’t invent this. It’s their version of a system called SynthID-Text, developed by Google DeepMind and published in the scientific journal Nature in 2024.



IN PRACTICE

Because the mark lives in the word choices and not in the file, it survives ordinary copying and pasting. Copy Claude's answer out of the app, paste it into an email or a Google Doc, and the pattern rides along with it. That's the part people are reacting to. It's also the part that has a much bigger catch than anyone's mentioning, which is section 4.

3 Images and files work completely differently

Text gets the word-choice pattern. Files get something else entirely, and it’s worth knowing the difference because their weaknesses aren’t the same.

When Claude generates an image or a file, formats like .png, .jpg, and .svg, it attaches signed provenance metadata using a standard called C2PA. Provenance just means origin story. It’s a record tucked into the file that says where this came from, what made it, and what’s been done to it since. It’s cryptographically signed, so tampering with it shows.

That sounds sturdier than a word pattern, and in one way it is. It’s much more specific. But it has a weakness the text mark doesn’t have: it’s attached to the file, not baked into the content. Anything that rewrites the file can knock it off.

IN PRACTICE

Converting the file to another format strips it. Re-saving it in an editor that doesn't understand C2PA strips it. Taking a screenshot creates a completely new file with no record at all. And this happens accidentally all the time: WhatsApp, iMessage, and Facebook all re-encode images when you upload them, which quietly removes the credentials. An image can lose its marking without anyone trying.

4 What it can't do (the part everyone's skipping)

This is where the online panic falls apart. The limitations aren’t secret and they aren’t in dispute. Anthropic has said them out loud.

Editing breaks it. Paraphrasing the text breaks it. Translating it breaks it. Mixing Claude’s writing with your own writing breaks it. Heavy editing breaks it. The pattern only holds up when the words stay mostly as Claude chose them, and the moment you

rework the sentences, you're replacing the very thing the mark is made of.

Short passages don't hold it. The mark is a statistical pattern, which means it needs enough words to measure. A paragraph or two might not carry enough signal to say anything with confidence. There's no exact word count where it flips on, it's a matter of degree, but short means unreliable.

A hit doesn't prove what people think. A positive detection means the content may have passed through Claude. It doesn't prove someone cheated, doesn't prove a human didn't write most of it, and doesn't prove the whole piece is AI-made. One AI-assisted paragraph in a long human-written document is a very different thing from an AI-written document, and the mark doesn't cleanly tell those apart.

A miss proves even less. No mark found doesn't mean a human wrote it. It could be from an older Claude model that isn't marked yet, from a different AI company entirely, or from Claude text that got edited. Absence of the mark is not evidence of anything.

And right now, nobody can check. Anthropic has said it will publish the technical details so people can detect its marks. Those tools are not public yet. So today, this moment, there is no tool you or a teacher or a client can run to test a piece of writing for Claude's watermark.

About that remover tool

If you came here from the video, here's exactly what that tool is and what each piece of it does for you.

Where to get it: github.com/guillaumemeyer/watermarks-remover

It's free, open-source, and it picked up more than 14,000 stars on GitHub in its first week. It's well built and it does three separate jobs, which is the part most coverage runs together.

Job one: invisible Unicode characters. Some AI tools slip in odd invisible characters, narrow spaces and similar, usually for formatting. Those are a genuine giveaway, they have nothing to do with the watermark in this guide, and stripping them is straightforward and reliable.

Job two: file metadata. It clears C2PA and EXIF data off images and documents. Also real, also reliable, and as section 3 covered, this is the same thing that happens by accident when you screenshot or re-save.

Job three: the actual word-choice watermark. This is the one everyone is talking about, and here's what the tool's own documentation says its method is: an AI rewrite. There's even a setting called paraphrase. It brings in a second AI to swap words for ones that mean the same thing, until the pattern breaks.

IN PRACTICE

So for the watermark itself, the tool's method is rewriting. That's worth knowing for two reasons. First, it's the same thing that happens when you edit AI writing into your own voice by hand, which costs nothing and takes no setup. Second, the tool's makers say plainly that they can't guarantee the result passes any given detector, and running your work through a cheaper AI to reword it can flatten your writing.

And there's one detail in the tool's own code worth knowing. It has a slot for a Claude text detector sitting there marked as a placeholder, waiting for Anthropic to ship one. The people who built the remover know perfectly well that nothing can read this mark yet.

So: genuinely useful for the invisible characters and the file metadata. For the word-choice watermark specifically, it's automating something you can already do by hand, with a second AI that may make your writing worse.

5 The part of this law that's actually about you

Here’s what almost no coverage mentions. Article 50 doesn’t only regulate AI companies. It also has rules for what the law calls deployers, meaning the people and businesses actually using AI systems and publishing the results. That’s you, if you use Claude for your work.

The good news is that the rules for deployers are narrower than people assume. There are two of them.

One: deepfakes have to be disclosed. If you publish AI-generated or manipulated image, audio, or video content that’s designed to look real, you have to say so. Think a visible label, an opening disclaimer, or a spoken note.

Two: AI text published to inform the public about matters of public interest has to be disclosed. That’s the specific phrase in the law. It’s aimed at news, current events, and public affairs, not at everything you post.

And there’s an exemption that matters more than either rule. If the AI-generated content went through real human review or editorial control, and an actual person or company holds editorial responsibility for it, the disclosure requirement doesn’t apply. The guidance is clear that this review has to be substantive. Skimming it and hitting publish doesn’t count.

IN PRACTICE

For most small businesses, this means less than you'd fear. Ordinary marketing content, product descriptions, your email newsletter, social captions, and website copy are not "informing the public on matters of public interest," so the text disclosure rule generally doesn't reach them. And if you're genuinely reading, editing, and standing behind what you publish, you're inside the editorial-control exemption anyway. Where you should pay attention: AI-made video or images of real-looking people or events, and any content you publish about news or public affairs.

6 If you're a student, read this part

A lot of the panic about this is coming from students, and it’s worth separating what’s actually true from what everyone assumes.

First: your school almost certainly isn’t using this. Schools run tools like Turnitin and GPTZero. Those are a completely different technology. They don’t read Claude’s watermark and never have. They guess, based on how the writing sounds, whether a machine wrote it. Claude’s mark is a separate system, and no one outside Anthropic can read it yet, including your school.

Second, and this is the part that matters most: the bigger risk to you is being wrongly accused, not being caught. The AI detectors schools already use are wrong more often than most teachers realize. One study of 500 human-written and 500 AI-written essays found the detector wrongly flagged human writing 4.2% of the time. A Stanford study found that for students who learned English as a second language, the false-positive rate jumped to 18.7%, meaning those students were nearly five times more likely to be wrongly accused.

These aren’t hypotheticals. A student lost a scholarship and landed on academic probation after using Grammarly to proofread a paper that got flagged as AI. An autistic student in the UK received a zero because their natural writing style read as machine-written to the software.

Third: a watermark hit wouldn’t prove what a school might think it proves. The mark says content may have passed through Claude. It can’t tell the difference between “Claude wrote my essay” and “I asked Claude to fix my commas.” Researchers covering this have flagged exactly that worry, that schools will treat a probabilistic signal as proof, and that nobody downstream will read the limitations.

IN PRACTICE

The single best protection is boring and it works: keep your drafts. Write in Google Docs or Word, where version history is saved automatically, and don't compose your paper in a chat window and paste it over. If your writing is ever questioned, a real revision trail showing the paper grow over days is far stronger evidence than any argument about detector accuracy. This protects you whether you used AI or not.

And the honest bottom line on the rules themselves: the EU law doesn’t decide what’s allowed in your class. Your school does. This watermark changes nothing about your school’s academic integrity policy, which was already the thing that governed what help you’re allowed to use. If you’re not sure where the line is for a specific assignment, ask your instructor before you turn it in, not after.

A few things to watch for

- This is a fast-moving area. Detection tools aren’t public yet, and when they arrive, the practical picture changes. What holds today is that nobody can check.
- Nothing here is legal advice, and nothing here is a ruling on your school’s policy. If you publish at real scale, especially in the EU, or you work in news or public affairs, talk to someone who does this for a living. If you’re a student, your instructor and

your school's integrity policy are the authority, not this guide.

- Other AI companies are heading the same direction. This isn't a Claude problem, and switching tools to avoid a watermark is a short-term move at best.
- The marking applies to newer Claude models now and older ones by December 2, 2026. If you're on an older model today, your output may not be marked at all yet.

The takeaway

Claude marks what it writes now, and that mark is real. But it was built to satisfy a transparency law, not to catch people, and it was never built to survive you rewriting the sentence.

So the fix isn't a tool. Read what AI writes, edit it into your own voice, and stand behind what you publish. That one habit breaks the watermark, keeps your writing good, and lands you inside the editorial-control exemption in section 5. It was already the right way to use AI. It just happens to also be the answer to this.

And if you're a student, the thing worth guarding against isn't this watermark. It's a detector your school already uses being wrong about you. Keep your drafts, know your school's actual policy, and ask before you turn something in rather than after.