M A V E R I C K A I

HOW TO REMOVE THE NEW CLAUDE WATERMARKS

THE FULL SETUP GUIDE

Claude just started watermarking everything it writes. Here is the setup in under five minutes, then what the watermark actually is and why the usual tricks do not touch it.

The full setup walkthrough, first thing in the guide

What the watermark really is (it is not hidden characters)

Why the EU is the reason this is happening now

The prompts that do the actual cleaning

The mistakes that make people think it did not work

What it can prove, and what nobody can prove yet

A one page quick reference at the end

2026 Edition

What Just Happened

On August 11 Anthropic announced that Claude is going to start watermarking the text it writes, and unlike most AI news this one affects you the moment you paste something into a doc. The reason it is happening now is the EU AI Act's Transparency Code, which came into force on August 2 and pushes the big labs to make AI generated content identifiable.

Aug 2

EU transparency rules came into force

Aug 11

Anthropic announced watermarking

5 min

to get the tool installed

Google already does this with Gemini, Anthropic is rolling it out now, and it is pretty safe to assume OpenAI and everyone else follows within the year. So this is not a Claude problem you can dodge by switching tools, it is where the whole industry is heading over the next twelve months.

The important thing to know before you start is that this is not the old invisible character trick where a zero width space gets slipped between your words. It is a completely different mechanic, which is why the advice you keep seeing in the comments does not work. There is a full explanation of how it works a few pages down, but the setup comes first because that is probably why you are here.

HEADS UP

Pasting your text into Notepad and copying it back out does nothing to this watermark. Neither does saving it as a different file type, or retyping the same sentences by hand. Those all used to work on the older marks, which is why the advice is still going around, but the new one lives somewhere else entirely.

Setting It Up

A developer called Guillaume Meyer published a free open source Claude skill called watermarks-remover, and it went round fast because it was out within days of the announcement. The skill inside it is called remove-ai-marks. This takes about two minutes and it is all point and click.

STEP 1: Open the repo

Go to **github.com/guillaumemeyer/watermarks-remover** in your browser.

STEP 2: Open the skills folder

In the file list near the top you will see a row labelled **skills/remove-ai-marks**. Click that.

STEP 3: Open SKILL.md

Inside you will see a file called **SKILL.md**. Click it to open it.

STEP 4: Download the raw file

In the top right of that file you will see **Download raw file**. Click it and the SKILL.md lands in your downloads.

STEP 5: Open your Claude skills

Go to Claude, click **Customize**, then **Skills**.

STEP 6: Upload it

Hit **Add** in the top right, choose **Upload skill**, then drag and drop the SKILL.md file straight in. That is the setup done.

TIP

Turn on code execution while you are in there, under Settings then Capabilities. The skill uses it to do the cleaning, and if it is off you will get a refusal that looks like the skill is broken when it is just a setting.

WANT THE FULL VERSION TOO?

The single file above gives you the text cleaning, which is what most people came for. If you also want the file tools, the ones that strip content credentials out of images and PDFs and audit a whole folder at once, grab the repo instead. Click the green Code button, then Download ZIP, unzip it, then zip up just the remove-ai-marks folder and upload that zip the same way. Same menu, same button.

Using It

Once it is installed you type the slash command and then paste your text or attach your file, and it runs an inspection first and tells you what it actually found before it changes anything.

THE COMMAND

```
/remove-ai-marks

Then paste your text underneath, or attach
the file you want cleaned.
```

Read the inspection report rather than skipping past it, because that is the part that tells you the truth. It gives you a count of the invisible characters it found and a list of any metadata worth stripping, and those numbers are real rather than estimated.

The rewrite pass

For prose it will then offer you the rewrite, and this is the step where you have to make a judgement call. A light pass keeps your writing intact but leaves more of the pattern behind, and a heavy pass scrambles the pattern properly but the writing comes back sounding like whichever model did the rewriting rather than sounding like you.

This is the paraphrase prompt the skill uses, if you want to run it yourself:

PARAPHRASE PASS

```
Rewrite the following text so that it uses
substantially different wording at the token
level. Change clause order, connectors, and
transition words; vary sentence boundaries and
length; and replace both content words and
function words where meaning allows. Preserve
all facts, numbers, names, and technical
identifiers. Do not add or remove claims.
Output only the rewritten text.

---
[paste your text here]
```

And this is the stronger one, which reads better but drifts further from what you originally wrote:

HUMANIZE PASS

```
Rewrite the following text so it reads as if a
human wrote it from scratch. Vary sentence
rhythm and length, replace formulaic AI-style
transitions and filler with concrete natural
phrasing, and use plain, varied wording.
Preserve all facts, numbers, names, and
technical identifiers. Do not add or remove
claims. Output only the rewritten text.

---
[paste your text here]
```

TIP

Run the invisible character clean again after any rewrite. The rewriting model is an AI too, so it will happily put a fresh set of odd characters back into your text on the way out.

What The Watermark Actually Is

Now the part worth understanding, because it is what makes the difference between using this well and using it badly.

When Claude writes a sentence there are usually several words that would fit equally well in any given spot, and the model is picking between them based on probability. The watermark works by nudging that choice using a secret key, so across a long enough passage the pattern of which near equivalent words got picked forms a signal that a detector holding the key can spot.

The text reads completely normally and no individual word looks unusual, because the signal only exists in the pattern across hundreds of words. There is nothing to find if you go looking at any one sentence.

> **THE SIMPLE VERSION**
>
> The watermark is not sitting in the file, it is sitting in the word choices themselves. Think of it less like a stamp on the paper and more like a particular handwriting style running through the whole page.

Where it applies

Anthropic has said it works at the model level rather than inside any one app, which means it is there whether the text came out of the Claude app, the API, Claude Code, or Cowork. Every model released after August 2 has it built in and they have said the older models are getting it too.

What actually weakens it

Because the signal is spread across a lot of words, the things that break it are the things that genuinely disturb that spread.

What you do	Effect on the watermark	Why
Fix a typo, tweak a few words	Barely touched	Most of the original word choices are still sitting there
Change headings and formatting	No effect	The signal is in the prose, not the layout
Rewrite heavily in your own words	Genuinely disrupted	You are replacing the choices that carried the pattern
Short text, a few sentences	Hard to read either way	Not enough words for the pattern to be reliable
Code	Weak to begin with	Little room to nudge word choice, and formatters strip the rest

What The Tool Actually Cleans

It works in three layers, and it is worth knowing which is which, because they are not equally reliable and the tool is upfront about that itself.

Layer	What it goes after	How reliable
A. Invisible characters	Zero width spaces, narrow spaces, bidirectional marks and other odd Unicode that AI tools slip in for formatting	Verifiable. It counts exactly what it pulled out and shows you
B. The statistical watermark	The word choice pattern, attacked by getting a second AI to rewrite your text	Best effort. Nobody outside the vendors can currently confirm it worked
Files	C2PA content credentials, EXIF and XMP metadata in PNG, JPEG, SVG, PDF, DOCX, ODT, HTML and MD	Verifiable. It reports each field it stripped

WHY LAYER A STILL MATTERS

Even though the new watermark is not made of invisible characters, AI tools do still leave those characters in your text, and they are a strong tell on their own. They also break things. Paste one into a CMS or a code editor and you get formatting bugs you will spend an hour hunting.

TIP

For Layer B the tool recommends rewriting with a different model than the one that wrote the text. So Claude text should not get rewritten by Claude, because you would just be swapping one model's marks for the same model's marks.

Where People Get This Wrong

These are the ones I keep seeing, and most of them are the difference between the tool doing something and the tool doing nothing while you think it worked.

Leaving code execution turned off

This is the number one reason people say it did not work. The skill installs fine and then cannot run anything, so you get a vague refusal instead of a report. Settings, then Capabilities, then turn it on.

Rewriting Claude text with Claude

If the model doing the rewrite is the same one that wrote the draft, you are handing the job to the thing you are trying to get away from. Use a different model for that pass, and a local one if you have the option.

Skipping straight past the inspection report

That report is the only genuinely verifiable output in the whole process. It tells you exactly what was found and removed, and everything after it is estimate rather than fact, so it is the one bit worth actually reading.

Running a light rewrite and calling it done

A gentle pass that keeps your voice intact is also a gentle pass that leaves most of the original word choices sitting there. You do not get both. Decide which one you actually want before you run it rather than hoping for the middle.

Forgetting to clean again afterwards

The rewriting model is an AI as well, so it puts its own crop of invisible characters in on the way out. Clean, rewrite, then clean again.

Forgetting that images and documents carry their own marks

People clean the text and then publish the screenshot next to it with full content credentials and camera metadata still attached. If the text mattered, run the files through as well, which is what the full version in the setup chapter is for.

> **TIP**
>
> Test it on something you have already published rather than on the thing you are about to send. You get to see what it actually finds without any pressure, and you will usually be surprised by how much metadata was riding along.

The Honest Part

I would rather you go in knowing the limits than trust this more than it deserves, so here it is straight.

The tool cannot prove it worked

The person who built it says so himself in the documentation. His line is that until vendors ship public detectors and keys, no tool can honestly certify that something fails the official check. Anthropic has announced a detection API but has not published how it scores, what its false positive rate is, or how much text it needs, so nobody outside Anthropic can currently verify whether a cleaned piece of text passes.

The invisible character removal is verifiable and the statistical layer honestly is not, and anyone telling you otherwise is guessing.

The rewrite costs you something real

Getting a second model to churn your word choices flattens tone and voice, and if the rewriting model is weaker than the one that wrote it you will feel that in the result. Your writing comes back a bit more generic every time you run a heavy pass, which matters a lot if the whole point was that it sounded like you.

Watermarks and detectors are two different problems

Plenty of AI detectors work on writing style rather than any hidden signal, so text can come out of this completely clean of watermarks and still get flagged by a tool that just thinks the writing reads like AI. Removing the watermark is not the same thing as passing a detector, and treating them as the same thing is how people end up surprised.

What is out of scope entirely

Pixel watermarks in images, audio and video marks, and C2PA soft binding, which is the kind that re-links to a remote record even after you strip the metadata. The tool is clear about all three.

Where This Is Genuinely Useful

The best uses for this are honestly the boring ones, and they are the reason I would keep it installed even if the watermark story went away tomorrow.

Use case	What it does for you
Pasting into a CMS or code editor	Clears the invisible characters that cause formatting bugs you would otherwise spend an hour hunting
Publishing images and PDFs	Strips EXIF, XMP and content credentials, which is basic privacy hygiene and something worth doing anyway
Auditing your own site	Runs across a folder or a sitemap and reports what your tools have quietly been leaving in your files
Cleaning your own drafts	Removes the small formatting tells before anything goes out under your name

And one last thought worth saying, because the developer who built this put it in his own documentation too. If you are writing in a situation where being caught using AI would actually be a problem, that is usually telling you something about the situation rather than handing you a problem to engineer around, and no amount of cleaning changes it. Use this on your own work, for your own reasons, and you will get real value out of it.

YOUR NEXT STEP

Grab the SKILL.md off GitHub, upload it under Customize then Skills, turn on code execution, and run it on something you have already published. Seeing what it finds sitting in your own old work is the fastest way to understand what this thing actually does.

Quick Reference

The setup in one block

```
SETUP
  # 1. Get the file
  github.com/guillaumemeyer/watermarks-remover
  Open skills/remove-ai-marks > SKILL.md
  Top right > Download raw file

  # 2. Upload it
  Claude > Customize > Skills
  Add (top right) > Upload skill
  Drag and drop SKILL.md

  # 3. Turn on code execution
  Settings > Capabilities > enable it

  # 4. Run it
  /remove-ai-marks then paste text or attach file
```

What to remember

Question	Short answer
Is it hidden characters?	No. It is the pattern of word choices across the whole passage
Does Notepad strip it?	No. Nothing about changing file type touches it
Does light editing strip it?	No. Most of the original word choices survive
What does weaken it?	Heavy rewriting, short text, and code
Which layer is provable?	Invisible characters and metadata. The rewrite is not
Does clean mean undetectable?	No. Style based detectors are a separate problem
Where does it apply?	Every Claude surface, because it works at the model level

SOURCES

Anthropic's announcement, August 11 2026. The EU AI Act Transparency Code, in force August 2 2026. The watermarks-remover repository and its README. Claude's help centre documentation on custom skills.

Want More AI Tips Like This?

I post daily AI content, tool breakdowns, and prompts that actually help you get more done. Follow me for the latest.

@MavGPT

TikTok | Instagram | Facebook

Clean your own work. Know what the tool can prove, and what it cannot.