AI STACK SIGNAL

CHECKED NOV 3, 2026

PLAIN-ENGLISH GUIDE · CLAUDE'S TEXT WATERMARK

CLAUDE NOW HIDES A WATERMARK IN ITS WORDS WHAT IT MEANS FOR YOU

Anthropic says text from new Claude models now carries an invisible watermark, worldwide. You can't see it, and right now you can't check for it yourself: the detector is a private preview for regulators, media, researchers, education groups and a few others. Inside: how it works in plain words, which Claude apps and models have it, what it does and doesn't do, who can apply for the detector, what it means for students, writers and businesses, myths vs facts, and every source.

01

No hidden characters

A secret key steers which of several good words Claude picks. That leaves a pattern only a detector with the key can spot.

02

Worldwide, every Claude app

Claude, Claude Code, Cowork, Claude Tag and the API. ChatGPT's watermark only covers the EU.

03

You can't check it (yet)

Detection is a private preview. Eligible organizations apply with a form.

!

A signal, not proof

Short text and code are hard to detect, a full rewrite wipes it, and no mark doesn't prove a human wrote it.

SOURCES: ANTHROPIC NEWS + HELP CENTER, TECHCRUNCH, THE DECODER, PCWORLD

01 START HERE
THE WATERMARK IN PLAIN WORDS

When Claude writes, it picks each next word from **several words that would all fit**. Normally a random number makes that choice. Now a **secret key** (plus the words before it) makes the choice instead. Over a long passage those choices form a statistical pattern. Anyone holding the key can test a passage for it. Nobody else can see it.

IN ANTHROPIC'S WORDS

The method "only changes the source of the randomness used to pick among words."

"Nothing is added to the text and there are no hidden characters."

It does not push Claude toward odd words. Anthropic's example: it would not make Claude pick an obscure word like "nubilous".

Aug 14
Anthropic's announcement, 2026 (updated Sep 1)

Aug 2
EU AI Act marking rules start, 2026

0
Hidden characters added to your text

0
Extra tokens, so no extra cost (Anthropic)

WHERE THE IDEA COMES FROM

- It's a version of **SynthID-Text**, the method Google DeepMind published in *Nature* in 2024.
- Part of a family of ideas going back to a 2022 proposal by researcher Scott Aaronson.
- OpenAI's own text watermark (textGrain) works on word choices too.

WHY NOW: THE EU AI ACT

From August 2, 2026 the EU requires AI providers serving its market to mark AI-generated content so other systems can identify it.

Anthropic signed the EU Code of Practice on Transparency of AI-Generated Content in July 2026, with around 190 signatories in total.

Files are different. Images and files Claude makes or processes (PNG, JPG, SVG and others) get a signed **C2PA Content Credential** in the file's metadata. Anyone can check those for free at **claude.com/check-files**. That checker does **not** read text.

02 WHERE THE MARK IS ON

APPS, MODELS, REGIONS

The watermark is added **at the model level**, so it doesn't matter which Claude app made the text. Anthropic applies it **worldwide**: "we don't yet have a durable way to scope it by region."

WHICH CLAUDE PRODUCTS

Claude (the apps), Claude Code, Claude Cowork, Claude Tag and the Claude Platform (API). Supported models on AWS, Google Cloud and Microsoft Foundry also carry text watermarks, per Anthropic's Help Center.

WHICH MODELS

New models mark text from day one. PCWorld reports Claude Fable 5.1 and Mythos 5.1 (Sep 1, 2026) were the first. Anthropic's Help Center table lists text watermarks for Fable 5.1, Fable 5, Mythos 5.1, Opus 5.5, Opus 5, Opus 4.8, Sonnet 5.5, Sonnet 5 and Haiku 5.5.

Older models: "Anthropic is adding watermarks to outputs from models released before August 2, 2026," rolled out "over the coming months". Text written before the mark existed carries no mark.

CLAUDE VS CHATGPT

	CLAUDE	CHATGPT (OPENAI)
Region	Worldwide	ChatGPT and Codex: EU only. API developers anywhere can switch it on for some models (off by default)
Started	Announced Aug 2026; first models Sep 1	Announced Oct 5, 2026
Name	(no product name; SynthID-Text based)	textGrain
Who can detect	Eligible organizations, private preview	Approved researchers and expert organizations
Public text checker	None	None

ChatGPT facts: TechCrunch, Oct 5, 2026. Claude facts: Anthropic news post and Help Center, PCWorld (Sep 1), TechCrunch (Aug 11).

03 HONEST LIMITS
WHAT IT DOES AND DOESN'T DO

A detection answers one question: **"What is the likelihood this was partly written by Claude?"** Anthropic is clear that this is **"a signal, not proof."**

WHAT ANTHROPIC SAYS IT DOES

- ✓ Travels with the text when it's copied and pasted.
- ✓ Probably survives light editing.
- ✓ Stays in translations Claude makes (Claude chose every word).
- ✓ Leaves quality, creativity and readability unchanged (Anthropic's internal tests; Google DeepMind found no rating difference on Gemini traffic).
- ✓ Costs nothing extra; speed impact "negligible".

WHAT IT DOESN'T DO

- ✗ It doesn't identify you: it "carries no identifying information and can't be traced to a specific person, organization, or chat."
- ✗ It can't tell "Claude wrote this" from "Claude heavily edited this".
- ✗ It can't prove a human wrote something.
- ✗ It can't detect other AI systems, even ones with their own watermarks.
- ✗ It doesn't change who owns the text or your rights under Anthropic's terms.

WHEN THE MARK IS WEAK OR MISSING

SITUATION	WHAT ANTHROPIC SAYS
Short text	Detection is weak on small samples; confidence rises with length.
Facts with one right answer	Few word choices, so few marks (example: "Mathematica" after "Principia").
Code and exact answers	Marks are sparse in code or "2 + 2 =" style outputs. Code comments can carry it.
Proofreading your own writing	Your words stay yours, so there's little for the mark to attach to.
Heavy edits or a full rewrite	A complete rewrite removes it; heavy paraphrasing or mixing can too.
Older Claude models	No mark until Anthropic adds support to them.

! **No numbers published.** Anthropic gives no accuracy rate, false-positive rate or minimum length for Claude's detector. Be careful with anyone quoting one.

04 THE DETECTOR
WHO CAN CHECK, AND HOW TO APPLY

Detection needs Anthropic's key, so it runs through Anthropic's **Detection API**. It is in **private preview**. PCWorld quotes Anthropic: the watermark "is invisible to anyone who does not have the detection API".

ELIGIBLE TO APPLY (ANTHROPIC'S LIST)

- ✓ Regulators
- ✓ Law enforcement
- ✓ Media and fact-checkers
- ✓ Independent researchers
- ✓ Educational organizations
- ✓ EU civil society groups
- ✓ Enterprises with their own verification duties under the EU AI Act

1 Check you fit the list

Apply as an organization. Being a member of the public isn't on the list.

2 Fill in Anthropic's form

The "Claude Watermark Detector Access Request Form", linked from Anthropic's Help Center and news post:

`forms.gle/9tGA33hPJJwtHsMk9`

3 Wait for access

Anthropic says it plans to expand access over time. It gives no timeline.

WHAT ANYONE CAN CHECK TODAY

Files only: claude.com/check-files reads the C2PA Content Credential in common image, video and audio files (up to 100 MB). It does not check text.

WATCH OUT FOR "CHECKER" SITES

Reading Claude's text watermark needs Anthropic's key. A website claiming to detect it for free isn't using Anthropic's detector, as far as any source we found shows. OUR TIP

05 OUR TAKE OUR TIPS WHAT IT MEANS FOR YOU

Facts on this page come from Anthropic. The advice is ours. The simple rule: **if Claude wrote it, say so where it matters.**



Students

Educational organizations can apply for the detector. Claude's words pasted into an essay can carry the mark, and light edits probably won't remove it. Follow your school's AI rules and say when you used Claude. Using Claude to proofread your own writing leaves your words yours.



Writers and freelancers

If a client's contract says "no AI text", Claude drafts can carry a mark. Agree on AI use up front. The mark carries nothing about who you are, so it can't name you or your account.



Businesses and marketers

Quality is unchanged, per Anthropic, so nothing changes in how Claude writes for you. If you publish AI content in the EU, check your own AI Act duties: Anthropic says builders should, and enterprises with verification duties can apply for the detector.



Teachers, editors, reviewers

A detection means Claude may have written or processed the text, not that the person cheated: Claude also proofreads, translates and summarizes. And no mark doesn't prove a human wrote it. Use it as one signal, then talk to the person.

✅ **Privacy check:** Anthropic says nothing in the watermark or its key reveals user information, and it can't be traced to a person, organization or chat.

06 QUICK ANSWERS
MYTHS VS FACTS

MYTH	FACT (SOURCE)
"It's invisible characters a text cleaner can strip."	No characters are added. The mark is in the word choices. (Anthropic)
"It makes Claude write worse."	Anthropic says it doesn't change quality, creativity or readability.
"It tells people who I am."	It carries no identifying information. (Anthropic, PCWorld)
"It's only in Europe."	Claude's is worldwide. ChatGPT's is EU-only in ChatGPT and Codex. (Anthropic, TechCrunch)
"Anyone can check my text now."	Text detection is a private preview for eligible organizations. (Anthropic)
"No watermark means a human wrote it."	It can't prove human authorship; old models, rewrites and short text leave no reliable mark. (Anthropic)

SOURCES (ALL READ NOV 3, 2026)

- Anthropic, "Claude text watermark" (Aug 14, 2026; updated Sep 1): https://www.anthropic.com/news/claude-text-watermark
- Claude Help Center, "How Claude marks AI-generated content": https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content
- Anthropic, "Check if a file was made with Claude": https://claude.com/check-files
- TechCrunch, Aug 11, 2026: https://techcrunch.com/2026/08/11/anthropic-says-it-will-watermark-text-generated-by-its-ai-models/
- The Decoder, Aug 14, 2026: https://the-decoder.com/anthropic-announces-watermark-detection-api-that-will-let-third-parties-detect-claudes-ai-texts/
- PCWorld, Sep 1, 2026: https://pcworld.com/article/3224834/claudes-updated-fable-and-mythos-models-watermark-their-replies.html
- TechCrunch, Oct 5, 2026 (OpenAI): https://techcrunch.com/2026/10/05/openai-will-start-watermarking-chatgpts-text-in-the-eu/

More plain-English AI breakdowns: aistacksignal.com

Not legal advice. Facts as published by the sources above on the date checked; Anthropic says detection access and model support will change over time. AI Stack Signal is not affiliated with Anthropic or OpenAI.