

■ // CLAUDE-WASSERZEICHEN · EU AI ACT

Claudes unsichtbares Wasserzeichen

Was wirklich markiert wird, ab welchem Modell, die 5 Wege, wie es zerbricht — und die False-Positive-Falle, die deinen eigenen Text erwischt.

WAS MARKIERT WIRD

AB MODELL 02.08.

DIE 5 LIMITS

FALSE-POSITIVE-FALLE

OHNE HYPE

Worum es geht: Seit dem 11. August webt Anthropic ein maschinenlesbares Wasserzeichen in Claudes Text-Output — wegen Artikel 50 des EU AI Act. Es ist unsichtbar, nicht abschaltbar und markiert unter Umständen auch Text, den **du** geschrieben hast. Hier steht, was wirklich passiert — und **wo die Grenzen liegen**.

1 Was genau passiert

Beim Generieren verzerrt Claude die Wortwahl minimal nach einem geheimen Schlüssel. Jede einzelne Wahl bleibt unauffällig — über einen längeren Text ergibt sich daraus eine **statistische Signatur**, die ein Detektor wiedererkennt (sog. Token-Selection-Bias: wo mehrere Wörter passen, wählt das Modell gesteuert).

- **Unsichtbar** — im fertigen Text nichts zu sehen
- **Nicht abschaltbar** — kein Schalter, kein Opt-out
- **Reist mit** — Copy-&-Paste in E-Mail, Doc oder Ticket trägt die Signatur weiter
- Dateien (SVG/PNG/JPG) bekommen zusätzlich signierte **C2PA**-Provenance-Metadaten

2 Ab welchem Modell — und wo

Betroffen sind alle Claude-Modelle, die **ab dem 2. August 2026** erscheinen (z. B. Opus 5). Ältere Modelle werden nachgerüstet. Das gilt in **Claude, der API, Claude Code und Cowork** — und zwar **weltweit**, nicht nur in der EU.

- Grund: EU AI Act, Artikel 50 (Transparenz KI-generierter Inhalte), Code of Practice Art. 50(2)
- Anwendbar ab 2. August 2026, volle Marktpflicht bis 2. Dezember 2026
- Anthropic, OpenAI, Google, Meta, Microsoft und Mistral haben den Kodex unterschrieben

3 Die 5 Grenzen — wo es zerbricht

Anthropic dokumentiert die Schwächen teils selbst. Das Wasserzeichen überlebt **nicht**:

- 1 **Starkes Umschreiben / Paraphrasieren** — wird die Wortwahl gründlich ersetzt, verschwindet die Signatur
- 2 **Übersetzen** in eine andere Sprache
- 3 **Kurze Texte** — auf wenigen Sätzen fehlt die statistische Signatur ganz
- 4 **Format-Konvertierung** (z. B. Code durch einen Formatter wie Prettier)
- 5 **Mischen von Mensch + KI** — eigener Text mit KI-Passagen verwässert das Muster

MERKSATZ

Anthropic selbst sagt, ein Treffer heißt nur „vielleicht mit Claude erstellt“ — die Hilfeseite formuliert es als „*may have been processed by Claude*“. Das ist ausdrücklich **kein Autorschaftsbeweis**.

4 Die False-Positive-Falle

Genau hier wird es absurd — und darum sind so viele sauer:

- **Dein eigener Text kann markiert werden**, wenn Claude ihn nur überarbeitet oder poliert hat. Die Signatur landet dann auf menschlicher Arbeit.
- **Kurzer KI-Text bleibt unmarkiert** — der eindeutige Fall fällt durchs Raster.

DER HAKEN

Das Feature markiert potenziell **die Falschen**: den ehrlichen Menschen, der Claude als Lektor nutzt — nicht den, der komplett generieren lässt.

5 Einordnung — warum jetzt

- **Nicht neu**: Google markiert Gemini-Text seit Mai 2024 mit SynthID.
- **Anthropic hatte 2024** einen ~99,9 % treffsicheren Text-Detektor gebaut — und wieder eingemottet, weil eine eigene Umfrage ergab, dass rund ein Drittel der Nutzer abspringen würde.
- **Mit-Motiv neben der EU**: Distillation-Angriffe chinesischer Labore (DeepSeek, Moonshot, MiniMax) — rund 24.000 Fake-Accounts, über 16 Mio. Anfragen, um Claude „abzukupfern“.

6 Was das für deine Inhalte heißt

SITUATION

WAS DU WISSEN MUSST

Panik?	Nein. Es ist statistisch, kein Beweis, und zerbricht leicht.
Ehrliche Nutzung	Wer KI transparent als Werkzeug einsetzt, hat nichts zu verbergen
Mensch + KI gemischt	Signatur ist ohnehin fragil — nicht als „Beweis“ oder „Schutz“ verlassen
Publisher / Agentur	Kennzeichnungspflichten für eigene Inhalte selbst prüfen (Art. 50)

■ // IN 5 SÄTZEN ERKLÄRT

Das Wichtigste auf einen Blick

Wenn dich jemand fragt — das reicht als Antwort.

- ✓ Claude verzerrt die Wortwahl nach einem geheimen Schlüssel — **unsichtbar, nicht abschaltbar**
- ✓ Ab Modellen vom **2. August**, weltweit, in Claude / API / Code / Cowork
- ✓ Zerbricht bei **Umschreiben, Übersetzen, kurzen Texten, Format-Wechsel, Mischen**
- ✓ Kann **deinen eigenen** (nur polierten) Text treffen — und kurzen KI-Text verfehlen
- ✓ Ein Treffer heißt **„vielleicht mit Claude“** — kein Autorschaftsbeweis

■ // BLEIB DRAN

KI-News **ohne Hype** — ehrlich eingeordnet, statt Panik-Schlagzeile.

Fandst du die Einordnung nützlich? Folg mir für die nächste —
verständlich erklärt, mit den echten Grenzen statt Clickbait.

@ki.borish

Fakten verifiziert am 14.08.2026 aus Anthropics Doku + öffentlicher Berichterstattung (The Register, The Verge, Daring Fireball) und einer 30-Tage-Recherche über Reddit (11.228 Upvotes), Hacker News (3.407 Punkte) und 25 YouTube-Erklärvideos. Dieser Guide ist eine technische Einordnung, keine Rechtsberatung — für verbindliche Kennzeichnungsfragen nach Art. 50 EU AI Act wende dich an einen Fachanwalt.