

GUIDA GRATUITA

Il marchio invisibile di Claude

Cosa dice davvero, cosa non lo cancella, e lo strumento nato in cinque ore per toglierlo

Cosa impone la legge, e da quando

L'articolo 50 dell'AI Act europeo si applica dal 2 agosto 2026. Da quel giorno chi costruisce sistemi che generano testo, immagini, audio o video deve marcarne le uscite in un formato leggibile a macchina.

Gira anche una seconda data, il 2 dicembre 2026, e viene raccontata spesso male. Non è la scadenza generale: è una proroga che vale soltanto per i sistemi già sul mercato prima del 2 agosto, e solo per l'obbligo di marcatura.

I modelli Claude usciti dal 2 agosto 2026 marcano dal primo giorno, perché per i sistemi nuovi la proroga non c'è. La parte obbligatoria è quella europea. Il marchio però Anthropic lo applica in tutto il mondo, e il motivo lo scrive: per ora non ha un modo stabile per limitarlo a una singola area. A luglio aveva anche firmato il codice di condotta europeo sulla trasparenza, che è volontario.

Chi deve dichiarare cosa

C'è poi la parte che riguarda chi pubblica, e va detta con precisione perché in giro viene raccontata molto più larga di com'è. Non riguarda qualsiasi testo scritto con l'AI.

Chi	Cosa deve fare
Chi costruisce il modello (Anthropic, OpenAI)	Marcare le uscite in formato leggibile a macchina
Chi pubblica un video o un'immagine manipolati	Dichiarare la manipolazione. Nelle opere dichiaratamente artistiche, satiriche o di finzione l'obbligo resta, ma basta segnalarlo in una forma che non rovini l'opera
Chi pubblica un testo fatto con l'AI	Dichiararlo solo se il testo informa su temi di interesse generale e nessuna persona lo ha riletto prima della pubblicazione

Quindi un preventivo per un cliente o un testo rivisto da te non ricadono in quell'obbligo. Ci ricade un articolo di informazione pubblicato senza che nessuno lo abbia controllato.

Cosa mette Claude, e dove

Dove	Che marchio	Si vede?
Nel testo	Un segno statistico dentro le parole scelte	No. Anthropic scrive che lettore e lettrice non distinguono un testo marcato da uno non marcato
Nei file (SVG, PNG, JPG)	Un dato di provenienza firmato, standard C2PA	No, sta nei dati del file. Sui file Anthropic non ha pubblicato una data netta come per il testo

Lo mette il modello, non la singola applicazione. Quindi vale sul sito, sull'app, e negli strumenti che si collegano a Claude.

Non è un simbolo nascosto nel testo. Quando Claude ha più parole equivalenti per dire la stessa cosa, sceglie seguendo uno schema segreto. Chi conosce lo schema, su un testo abbastanza lungo, riconosce che quella sequenza di scelte non è casuale. Non c'è niente che tu possa cercare o cancellare a mano.

IL PUNTO CHE QUASI NESSUNO HA LETTO

Quel marchio dice una cosa sola, e nemmeno con certezza: probabilmente Claude ha messo mano a questo testo, a un certo punto. Non dice chi l'ha pensato, non dice quanto ci hai messo mano. Anthropic stessa scrive che non è una prova definitiva.

La buona notizia, se correggi soltanto

La buona notizia, se correggi soltanto

Anthropic scrive che il marchio si attacca solo alle parole scelte da Claude. Se il testo l'hai scritto tu e gli hai fatto solo una correzione, le parole restano quasi tutte tue, e il marchio non ha quasi niente a cui attaccarsi. Il limite vero è un altro: dove il marchio c'è, non distingue un testo scritto da Claude da uno che Claude ha riscritto pesantemente.

Cosa lo cancella, e cosa no

Copiare e incollare non basta

È la confusione più diffusa. Il segno nel testo viaggia insieme alle parole: se selezioni, copi e incolli altrove, resta. Sopravvive anche a modifiche leggere.

L'elenco che segue non è una critica di qualcuno: è scritto nella pagina di assistenza di Anthropic. Il marchio diventa non rilevabile quando succede una di queste cose.

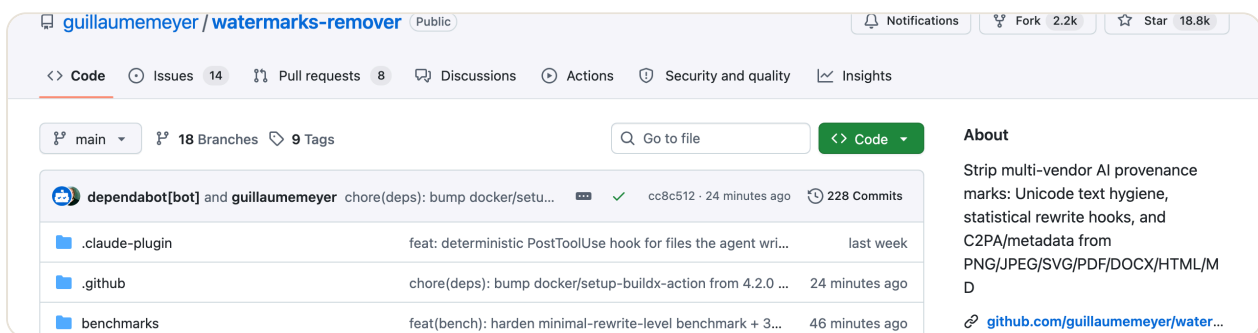
- Il testo viene riscritto pesantemente, parafrasato o tradotto
- Il passaggio è troppo corto perché il segno regga
- Il file viene convertito in un altro formato e i dati di provenienza si perdono
- Del contenuto viene fatto uno screenshot
- La piattaforma dove pubblichi non gestisce quel tipo di marcatura

Lo screenshot rende l'idea. Basta una fotografia dello schermo e la provenienza dell'immagine sparisce, senza niente di sofisticato.

Cinque ore

Anthropic annuncia il marchio l'11 agosto. Guillaume Meyer, che a Parigi ha fondato Memo, una società che fa strumenti di AI per i negozi online, pubblica su GitHub la prima versione di uno strumento per toglierlo. Le circa cinque ore di lavoro sono una cosa che ha dichiarato lui alla stampa, non un dato che si legge sul progetto.

Si chiama watermarks-remover. In due giorni aveva superato le quattromilacinquecento stelle su GitHub, il sito dove i programmatori pubblicano il codice, e il 27 agosto era arrivato a diciottomilasettecento.



La pagina del progetto il 27 agosto 2026: 18.700 stelle e 2.200 copie.

Cosa fa, in concreto

- Toglie dal testo i caratteri invisibili e le spaziature anomale
- Riscrive il testo per disturbare il segno statistico, e l'autore lo definisce un tentativo, non una garanzia
- Ripulisce i dati di provenienza dei file: C2PA, EXIF, XMP e le proprietà dei documenti

Da non confondere

I caratteri invisibili non sono il marchio di Claude, sono un'altra cosa. Toglierli non tocca il segno statistico dentro le parole.

Esiste anche in versione da attivare dentro Claude come competenza. Il pacchetto pronto però non si scarica: va generato lanciando uno script dal progetto, oppure installato con un comando da sviluppatori. Non ti spiego passo passo come si fa, e alla pagina dopo trovi il perché.

Il problema che ribalta la storia

Nessuno può dimostrare che funzioni davvero sul marchio dentro il testo.

Anthropic non ha ancora pubblicato lo strumento per leggerlo. Senza quello non esiste modo di verificare se un documento ripulito porti ancora il segno oppure no.

Lo scrive l'autore stesso

Nel progetto è dichiarato che nessuno strumento può onestamente certificare di aver superato il controllo ufficiale, e che l'ultima parola resta a chi il marchio lo ha messo.

Quindi va separato quello che si vede da quello che si promette. La parte sui caratteri nascosti e sui dati dei file è verificabile: apri il file e guardi. Quella sul testo, per adesso, è una promessa che nessuno può misurare.

Non è un caso isolato. Il 13 agosto, due giorni dopo l'annuncio, il sito di sicurezza informatica BleepingComputer aveva già censito diversi progetti su GitHub e alcuni siti nati apposta, e concludeva che quasi nessuno riesce a dimostrare quello che promette.

Cosa cambia per te, oggi

Oggi, in pratica, niente: lo strumento per leggere il marchio non è ancora pubblico, quindi nessun cliente può darlo in pasto a un programma di verifica e ottenere una risposta. Anthropic scrive che rilascerà presto un servizio per rilevarlo, e il marchio resta dentro i testi che scrivi adesso.

Una cosa da non confondere: i rilevatori di testo AI che girano da anni esistono già e qualcuno li usa. Sbagliano spesso, non c'entrano niente con questo marchio, e non spariscono perché il marchio non è leggibile.

Quello che cambia è quando conviene decidere la tua regola: adesso, con calma, invece che dopo sotto pressione.

Fai così

- ✓ Metti per iscritto cosa fai passare dall'AI e cosa no, se consegni testi a clienti
- ✓ Rileggi e correggi sempre quello che consegni, a prescindere dal marchio
- ✓ Ricorda che uno screenshot di un'immagine ne cancella la provenienza

Evita

- ✗ Dare per scontato che il marchio dimostri chi ha scritto il testo
- ✗ Fidarti di strumenti che promettono di superare un controllo che non esiste ancora
- ✗ Confondere la proroga di dicembre con un rinvio generale

DIECI MINUTI, UNA VOLTA SOLA

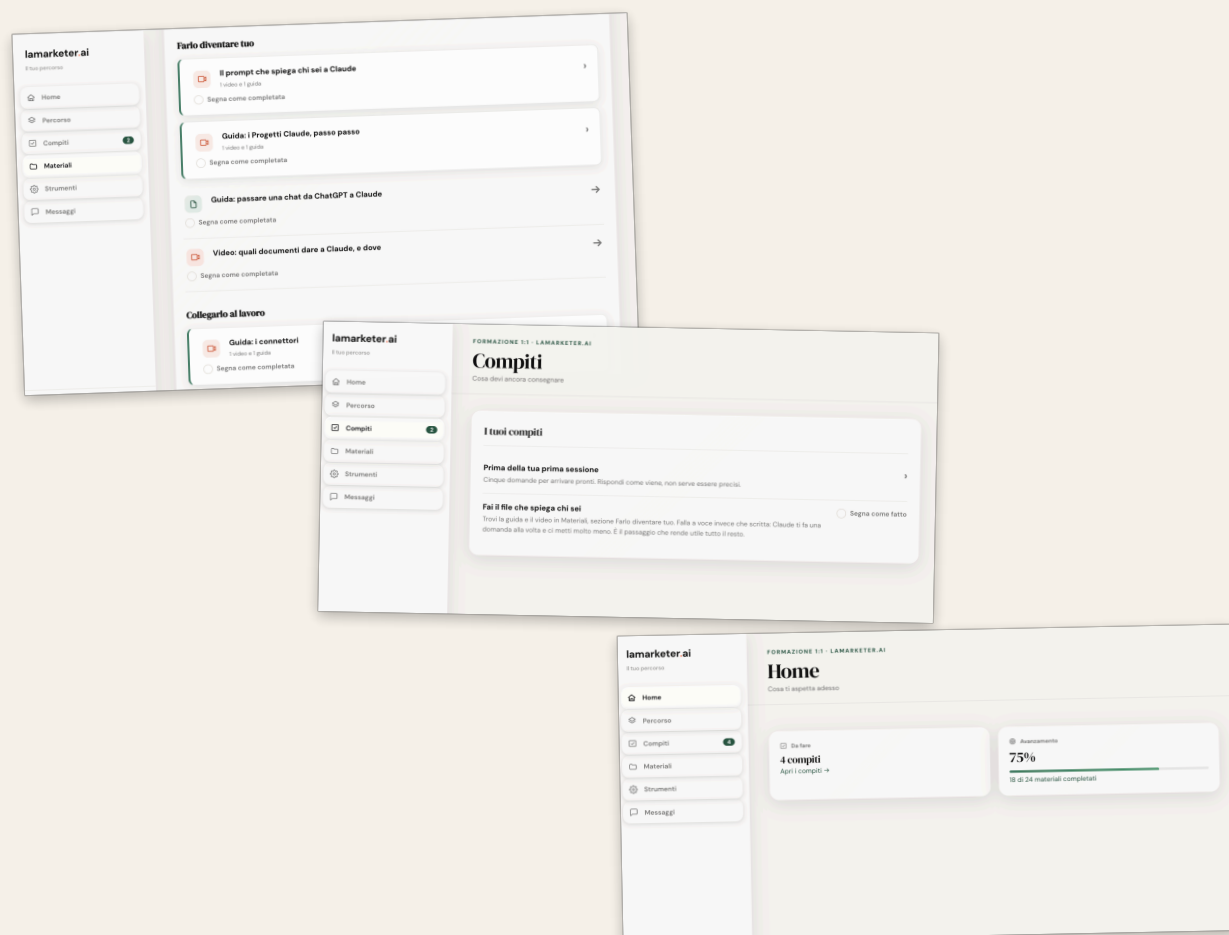
Scrivi la tua regola in tre righe e tienila dove la ritrovi. Serve a te prima che ai clienti.

1. Elenca i tipi di testo che consegni: preventivi, relazioni, post, documenti per i clienti.
2. Per ognuno scrivi cosa fa l'AI: niente, correggo, scrivo insieme, scrive e io rivedo.
3. Decidi quali di questi dichiarare al cliente, e in che forma. Una riga nel preventivo basta.

Non è un obbligo per te nella maggior parte dei casi. È il modo di non farti trovare impreparata da un cliente che te lo chiede, che succede molto prima di qualsiasi legge.

Cosa significa lavorare con me

Il percorso 1:1 non è solo le sessioni live. Dentro la piattaforma trovi i materiali costruiti sul tuo caso, i compiti tra una sessione e l'altra e l'avanzamento reale, così tra un incontro e il successivo hai sempre dove tornare.



Materiali, compiti tra una sessione e l'altra, e l'avanzamento reale del percorso.

IL PASSO SUCCESSIVO

Il percorso 1:1.

5 sessioni costruite intorno ai tuoi obiettivi professionali e personali.

5 posti al mese. Prima guardiamo il tuo caso, poi ti dico se posso aiutarti.

Prenota la call gratuita →