

KI-Wasserzeichen: Was der Clip belegt

Die vollständige Entfernung ist nicht nachgewiesen. Unsichtbare Zeichen, Dateimetadaten und statistische Textmarkierungen müssen getrennt geprüft werden.

Die beiden eingereichten Videos sind jeweils rund 33,7 Sekunden lang und zeigen nach Bild- und Untertitelabgleich denselben Beitrag. Er bewirbt ein Werkzeug zum Entfernen von KI-Markierungen und nennt dabei ChatGPT, Claude, Astra 6 und Fable 5.1.

Aussagen im Quellenabgleich

Aussage / Beobachtung	Bewertung
KI markiert Texte und Dateien Im Clip etwa 00:05-00:11	Nur mit genauer Angabe des Verfahrens. Anthropic beschreibt statistische Textmarkierung. OpenAI dokumentiert C2PA und SynthID für Bilder sowie SynthID für unterstütztes Audio. Diese OpenAI-Quelle bestätigt keine pauschale Textmarkierung durch Astra 6. [1] [2]
Zeichenbereinigung und Claude-Wasserzeichen Technische Einordnung	Verschiedene Verfahren. Laut Anthropic liegt das Signal in der Wortauswahl; zusätzliche unsichtbare Zeichen werden dafür nicht eingefügt. Eine reine Zeichenbereinigung belegt deshalb keine Entfernung dieses Signals. [1]
Werkzeug liefert eine bereinigte Kopie Vorführung etwa 00:14-00:24	Bedienung gezeigt, vollständige Entfernung unbelegt. Die sichtbare Wortfolge bleibt gleich. Ein geeigneter Wasserzeichentest vor und nach der Bearbeitung fehlt. Auch die Anbieterin garantiert keine vollständige Entfernung statistischer Textmarkierungen. [4]

Was in der Vorführung fehlt

Um 00:14 steht der sichtbare Schalter „Claude Rewriter“ auf aus. Um 00:18 wird ein kurzer Text eingefügt; anschließend erscheint eine Erfolgsmeldung. Es gibt keinen Nachweis, dass der Ausgangstext überhaupt ein auswertbares Wasserzeichen enthielt. Die Meldung allein zeigt auch nicht, welche Zeichen oder Metadaten verändert wurden.

Die Werkzeugseite unterscheidet selbst Zeichenbereinigung, Metadatenentfernung und optionales Umschreiben. Eine umfassende Erfolgsgarantie lässt sich daraus nicht ableiten. [3]

Prüfumfang: sichtbare Videoinhalte, eingeblendete Untertitel und die verlinkten Quellen. Kein Abhören oder Transkribieren der Tonspur; kein eigener Funktionstest des Onlinedienstes. Zeitangaben sind Näherungswerte.

Vier verschiedene Signale

Die Bezeichnung „Wasserzeichen“ reicht für eine technische Bewertung nicht aus.

Signal	Was es bedeutet	Grenze der Aussage
Unsichtbare Unicode-Zeichen	Unicode enthält reguläre Steuerzeichen etwa für Schreibrichtung und Zeichenverbindungen. [6]	Ihr Vorhandensein beweist keine KI-Herkunft. Blindes Löschen kann reguläre Textfunktionen verändern.
Statistisches Textwasserzeichen	Bei SynthID-Text verändert sich die Auswahl von Tokens während der Erzeugung. Ein passender Prüfer sucht das daraus entstehende Muster. [5]	Ein Zeichensäuberer weist die Entfernung dieses Musters nicht nach. Kurze Texte liefern oft zu wenig Signal für eine verlässliche Prüfung. [1]
Dateimetadaten / Content Credentials	C2PA kann signierte Herkunfts- und Bearbeitungsangaben mit einer Datei verbinden. Solche Metadaten können verloren gehen oder entfernt werden. [2]	Fehlende Metadaten beweisen keine menschliche Urheberschaft. Metadatenbereinigung sagt nichts über zusätzliche Signale im Inhalt aus.
Eingebettetes Signal in Bild oder Audio	OpenAI beschreibt SynthID als zusätzliche Ebene zur Herkunftskennzeichnung unterstützter Medien. [2]	Die Blotato-Einbindung des Sabrina-Cleaners schließt die Entfernung von OpenAI-SynthID ausdrücklich aus. [7]

Herkunft, Urheberschaft und Datenschutz

Anthropic zufolge kann das Textsignal eine Beteiligung von Claude anzeigen, aber Erzeugung und starke Überarbeitung nicht auseinanderhalten. Es enthält laut Anbieter keine Kennung einer Person, Organisation oder Unterhaltung. Das ist eine Herstellerangabe, keine hier durchgeführte unabhängige Prüfung. [1]

Ein allgemeiner KI-Detektor ist wiederum etwas anderes: Er kann Texte anhand gelernter Merkmale einstufen, ohne ein eingebettetes Wasserzeichen auszulesen. Die SynthID-Studie unterscheidet diese Ansätze und beschreibt Fehlklassifikationen solcher allgemeinen Detektoren. [5]

Vorgeschlagener Maßstab für AIRW.org

Nachweis verlangen: Modell und Version, markierten Ausgangstext, Bearbeitung sowie passenden Prüfer dokumentieren. Unmarkierte Vergleichstexte und mehrere Proben helfen, Zufallstreffer zu erkennen. Eine grüne Erfolgsmeldung ersetzt diese Prüfung nicht.

Datenwege klären: Die Blotato-Einbindung beschreibt eine externe Verarbeitung durch Sabrinas Dienst. Das ist kein Nachweis lokaler Verarbeitung. Die Datenwege der im Clip gezeigten Direktseite wurden hier nicht technisch untersucht. [7]

Beitragsentwurf für AIRW.org

Vertrauen braucht überprüfbare Herkunft

Ein kurzer Clip verspricht, KI-Wasserzeichen aus Texten und Dateien zu entfernen. Seine Vorführung zeigt einen eingefügten Text und eine bereinigte Ausgabe. Ob vorher ein Wasserzeichen vorhanden war und welches Signal nachher verschwunden ist, bleibt offen.

Die technischen Unterschiede entscheiden über die Aussagekraft. Unsichtbare Steuerzeichen, Dateimetadaten und Muster in der Wortwahl erfüllen verschiedene Aufgaben. Ein Verfahren, das eine dieser Ebenen verändert, muss deshalb noch keine andere beeinflussen. [2] [5] [6]

Für AIRW.org ergibt sich daraus eine Frage der Selbstbestimmung: Menschen sollten nachvollziehen können, welche Informationen ein Werkzeug verarbeitet, welche Spuren es erzeugt und welche Behauptungen sich nachprüfen lassen. Anbieter sollten Prüfverfahren und Fehlergrenzen offenlegen. Wer eine Entfernung verspricht, sollte einen nachvollziehbaren Vergleich vorlegen.

Ebenso braucht die Bewertung eines Beitrags mehr als ein Herkunftssignal. Quellen, sachliche Richtigkeit, menschliche Bearbeitung und redaktionelle Verantwortung gehören zur Prüfung. Ein Text wird durch eine bereinigte Datei weder wahr noch falsch. Die Verantwortung für seine Veröffentlichung bleibt eine eigenständige Aufgabe.

Vorgeschlagene Leitlinie: Herkunft transparent machen, Privatsphäre schützen und technische Versprechen an überprüfbaren Ergebnissen messen.

Quellen und Grenzen

Abruf: 17.09.2026. [1]-[2] Herstellerangaben; [3]-[4] und [7] Selbstauskünfte der Werkzeuganbieter; [5] Forschungsarbeit zum Verfahren, kein Test dieses Clips; [6] technische Unicode-Dokumentation.

[1] Anthropic: How Claude's text watermark works (14.08.2026)

[2] OpenAI: Advancing content provenance (19.05.; aktualisiert 31.07.2026)

[3] Sabrina Tools: Watermark Remover, Funktionsbeschreibung

[4] Sabrina Ramonov: How to make ChatGPT & Claude sound like a human (21.08.2026)

[5] Dathathri et al.: Scalable watermarking for identifying large language model outputs, Nature (23.10.2024)

[6] Unicode Consortium: Display of Unsupported Characters

[7] Blotato: AI watermark remover, Leistungsumfang und Datenverarbeitung

Videogrundlage: savetik-hd(20260917-181950).mp4 und savetik-hd(20260917-181959).mp4. Zwei inhaltlich gleiche Einsendungen sind keine unabhängigen Bestätigungen. Dieser Text ist ein redaktioneller Vorschlag für AIRW.org.