Kundeninformation

KI-Wasserzeichen und Content-Kennzeichnung

Was Sie über KI-generierte Inhalte wissen sollten – und wie wir bei K12 damit umgehen

Was ist ein KI-Wasserzeichen? Anthropic, Google und andere Anbieter versehen KI-generierte Texte und Bilder mit unsichtbaren Signalen, die nachweisen sollen, dass ein Modell daran beteiligt war. Für uns als Kommunikator:innen ist das kein abstraktes Thema, sondern eines, das unsere tägliche Arbeit betrifft.

Transparenz über KI-Einsatz ist richtig, der Kampf gegen billig produzierten „KI-Slop" ebenso. Nur trifft die technische Umsetzung dabei unscharf: Wer einen Text selbst schreibt und KI nur zur Korrektur oder Strukturierung nutzt, kann unter denselben Generalverdacht geraten wie eine komplett KI-generierte Textmasse. Genau diesen Unterschied halten wir in diesem Dokument fest – und wir beobachten das Thema weiter, weil es sich schnell entwickelt.

Zusammenfassung der wichtigsten Punkte

1. Ein Wasserzeichen kann auch da auftauchen, wo Idee, Konzept und Text größtenteils von uns sind. Es belegt Verarbeitung, nicht Autorschaft. Wer selbst konzipiert, schreibt und KI nur unterstützend einsetzt, kann trotzdem ein Signal erzeugen – ohne dass das etwas über die eigentliche Urheberschaft aussagt.
2. Aktuell ist für Außenstehende kaum überprüfbar, ob ein Text ein Wasserzeichen trägt. Das gilt in beide Richtungen: Auch die Behauptung, ein fremder Text sei „garantiert KI-generiert", ist auf absehbare Zeit nicht seriös belegbar.
3. Es ist möglich, dass entsprechende Prüf-Tools (etwa das von Anthropic angekündigte) künftig leichter zugänglich werden. Dann würde ein Wasserzeichen leichter sichtbar.
4. Gründliche redaktionelle Überarbeitung ist nicht nur handwerklich richtig, sondern verändert auch die Wahrscheinlichkeit, dass ein Wasserzeichen „mitgeschleppt" wird.
5. Bei kurzen Texten sind Wasserzeichen in der Regel nicht enthalten – der normale Social-Media-Post ist davon praktisch nicht betroffen.
6. Website-Texte prüfen wir vorsorglich zusätzlich, bevor sie veröffentlicht werden. Ein künftiger Effekt auf die Sichtbarkeit in Suchmaschinen ist nicht auszuschließen, auch wenn er aktuell nicht bestätigt ist.
7. Wichtig festzuhalten: Entfernte Metadaten oder Wasserzeichen beweisen nicht, dass ein Inhalt menschlichen Ursprungs ist – sie beweisen nur, dass die Aufzeichnung fehlt. Umgekehrt ist ein Wasserzeichen bei Texten kein Beweis dafür, dass ein Text überwiegend KI-generiert ist.
8. Unabhängig davon: Wir machen KI-Einsatz Ihnen gegenüber transparent.

1 Worum es geht

Im Zuge des EU AI Act verpflichten sich große KI-Anbieter, KI-generierte Inhalte technisch nachvollziehbar zu machen. Anthropic (Claude), Google (Gemini) und andere Anbieter setzen dafür zwei Verfahren ein: ein unsichtbares Muster in der Wortwahl bei Texten sowie signierte Herkunftsdaten (C2PA) bei Bildern und Dateien. Beide Mechanismen zeigen ausschließlich, dass ein Modell an der Entstehung beteiligt war – nicht, wer den Inhalt inhaltlich verantwortet hat.

2 Wie funktioniert die Kennzeichnung?

Text: Ein Sprachmodell wählt an vielen Stellen zwischen mehreren gleichwertigen Wörtern. Ein geheimer, nur dem Anbieter bekannter Schlüssel verschiebt dabei leicht, welches Wort bevorzugt gewählt wird. Über einen längeren Text ergibt das ein für Menschen unsichtbares, aber statistisch messbares Muster – der Inhalt und die Aussage des Textes ändern sich dadurch nicht.

Bilder und Dateien: Bei Bildern und Dateien hängen die Anbieter signierte Metadaten nach dem offenen Standard C2PA an, den auch Kameras und etablierte Bildbearbeitungssoftware seit Jahren zur Herkunftsdokumentation nutzen. Google und teilweise OpenAI ergänzen das bei Bildern zusätzlich um ein robusteres, direkt in die Pixel eingebettetes Wasserzeichen (SynthID).

3 Was sehen Sie als Leser:in davon?

Nichts. Es gibt kein Icon, keinen Hinweistext und keine wahrnehmbare Veränderung an Text oder Bild – die Kennzeichnung läuft vollständig im Hintergrund. **Das darf nicht mit der gesetzlichen Pflicht verwechselt werden, KI-Einsatz bei Veröffentlichungen zu einem Thema von öffentlichem Interesse sichtbar offenzulegen:** Diese Pflicht besteht unabhängig von der unsichtbaren Markierung, betrifft aber ohnehin nur einen kleinen Teil der Inhalte, die wir für Sie erstellen.

4 Lässt sich das nachweisen?

Für Außenstehende praktisch nicht. Keiner der großen Anbieter bietet aktuell ein frei zugängliches Werkzeug, mit dem sich ein Text zuverlässig auf ein Wasserzeichen prüfen lässt. Bei Bildern ist die Lage etwas besser: Signierte C2PA-Metadaten lassen sich mit gängigen Standard-Tools auslesen.

Von einfachen Online-„KI-Detektoren" raten wir ausdrücklich ab. Sie prüfen kein echtes Wasserzeichen, sondern raten anhand von Sprachmustern – und liegen nachweislich häufig falsch, auch bei Texten, die vollständig von Menschen geschrieben wurden.

5 Risiken im Überblick

Ein Wasserzeichen zeigt Verarbeitung, nicht Autorschaft: Auch Texte, die überwiegend von Menschen stammen und nur redigiert wurden, können ein Signal enthalten. Genau deshalb prüfen und verantworten wir jeden Inhalt inhaltlich selbst, unabhängig davon, ob und wie KI beteiligt war.

Stilbasierte „KI-Detektoren" – nicht zu verwechseln mit einer echten Wasserzeichen-Prüfung – sind nachweislich fehleranfällig und können auch rein menschliche Texte fälschlich verdächtigen.

Suchmaschinen-Ranking: Google erklärt öffentlich und wiederholt, dass die Entstehungsart eines Textes – Mensch oder KI – für sich genommen kein Rankingfaktor ist; bewertet werden Qualität, Nützlichkeit und erkennbare eigene

Erfahrung. Eine offizielle „KI-Text-Abwertung" gibt es **nach aktuellem Stand** nicht, und ein Wasserzeichen spielt dabei bislang nachweislich keine Rolle als Rankingsignal.

Ein praktisches, indirektes Risiko bleibt trotzdem: Massenhaft produzierte, inhaltsarme KI-Texte ohne eigene Substanz fallen eher durch die ohnehin bestehenden Qualitätsfilter. Das betrifft aber die Qualität des Inhalts, nicht die Tatsache, dass KI beteiligt war – und ist der Grund, warum wir jeden Text vor Veröffentlichung inhaltlich prüfen und schärfen, statt uns auf die reine KI-Ausgabe zu verlassen. Wir beobachten diesen Punkt weiter, da sich die Praxis der Suchmaschinen ändern kann.

Kurze Formate wie Social-Media-Posts sind von alldem praktisch nicht betroffen, da sie zu wenig Textmaterial für ein messbares Signal enthalten.

6 Lässt sich eine Kennzeichnung entfernen?

Bei Text lässt sich das nicht garantieren: Niemand kann mit Sicherheit sagen, dass ein Text danach „sauber" ist. Eine gründliche redaktionelle Überarbeitung – eigene Struktur, eigene Wortwahl, keine wortgleich übernommenen Passagen – verändert aber automatisch auch das technische Muster im Text und senkt damit die Wahrscheinlichkeit, dass ein Wasserzeichen erhalten bleibt, spürbar. Das ist bei uns ein Nebeneffekt sauberer redaktioneller Arbeit, kein gezielter Trick.

Bei Bildern ist die Lage eindeutiger: Einfache Metadaten lassen sich technisch entfernen. Das beweist dann aber nicht, dass ein Bild menschlichen Ursprungs ist, sondern lediglich, dass die Aufzeichnung fehlt.

7 Welche Plattformen sind betroffen?

Anbieter	Text	Bild / Datei
Claude (Anthropic)	Unsichtbares Wortwahl-Wasser-zeichen, seit 2. August 2026 aktiv.	Signierte C2PA-Metadaten, kein Pixel-Wasserzeichen.
Gemini (Google)	SynthID-Text, bereits seit 2024 aktiv.	SynthID-Pixel-Wasserzeichen, robust gegen Zuschnitt und Kompression, zusätzlich C2PA.
ChatGPT (OpenAI)	Seit Oktober 2026 aktiv.	C2PA-Metadaten, teils zusätzlich SynthID.
Copilot (Microsoft)	Kein eigenes Wasserzeichen; hängt vom jeweils antwortenden Modell im Hintergrund ab.	Signierte C2PA-Metadaten („Content Credentials").

Der Stand entwickelt sich laufend. Wir prüfen die Angaben regelmäßig neu.

Was das für die Zusammenarbeit mit uns bedeutet

Wir setzen KI in unseren Workflows gezielt und transparent ein – als Werkzeug für Recherche, Struktur und Tempo, nicht als Ersatz für redaktionelle Verantwortung. Inhalte, für die wir stehen, werden von uns konzipiert, geprüft und verantwortet. Das Thema Kennzeichnung entwickelt sich schnell; wir beobachten es laufend für Sie und melden uns, sobald sich etwas ändert, das für Ihre Inhalte relevant ist.

Stand: 14.10.2026. Bei Fragen zu einzelnen Inhalten sprechen Sie uns gerne an.