

Unsichtbare Wasserzeichen bei Claude

Unsichtbare Wasserzeichen in Texten, die Claude generiert oder bearbeitet hat – Was hierzu bislang bekannt ist

Anthropic hat angekündigt, dass neue Claude-Modelle künftig KI-generierte Texte maschinenlesbar markieren. Hierfür soll ein unsichtbares Wasserzeichen direkt in den Text eingebettet werden.

Soweit diese Modelle Bestandteil eines KI-Systems sind, mit dem KI-generierte Inhalte erzeugt werden, soll damit die Transparenzpflicht aus Art. 50 Abs. 2 KI-VO erfüllt werden. Davon zu unterscheiden ist die Kennzeichnungspflicht des Betreibers eines KI-Systems nach Art. 50 Abs. 4 KI-VO.

1 Wie funktioniert das Wasserzeichen?

Das Wasserzeichen entsteht durch ein statistisches Muster bei der Auswahl der Wörter beziehungsweise Tokens. Es werden dabei keine versteckten Zeichen oder besonderen Formatierungen in den Text eingebaut. Das Verfahren beruht vielmehr auf der Wortfolge des erzeugten Textes.

Dieses Verfahren ist nicht neu und ähnelt Verfahren wie SynthID-Text von Google.

Wichtig

Das Verfahren kann nicht nur bei vollständig neu generierten Texten zum Einsatz kommen. Auch bei Aufgaben wie Korrekturen, Übersetzungen oder Zusammenfassungen kann ein entsprechendes Muster entstehen, soweit Claude selbst Wörter für den Text auswählt.

2 Was zeigt das Wasserzeichen genau?

Es zeigt mit einer bestimmten Wahrscheinlichkeit, dass Claude an einem Text beteiligt war.

Es lässt sich daraus jedoch nicht ablesen, ob Claude den Text vollständig erstellt, einzelne Passagen ergänzt oder einen bereits vorhandenen Text überarbeitet hat. Das Wasserzeichen gibt daher keinen Aufschluss über den Umfang der KI-Nutzung.

3 Bleibt das Wasserzeichen beim Kopieren erhalten?

Ja. Beim Kopieren und Einfügen des Textes bleibt das Wasserzeichen erhalten, da sich die Wortfolge dabei nicht verändert.

4 Verschwindet das Wasserzeichen beim Umschreiben?

Das Wasserzeichen kann bei einer nachträglichen Veränderung des Textes verloren gehen. Anthropic nennt selbst verschiedene Vorgänge, bei denen dies der Fall sein kann, etwa echtes Umschreiben und Paraphrasieren, Übersetzen, starkes redaktionelles Überarbeiten oder das Vermischen mit eigenem Text.

In diesen Fällen werden Wörter neu gewählt bzw. verändert, sodass das ursprüngliche Muster nicht notwendigerweise erhalten bleibt.

Wichtig

Wo im Einzelfall die Grenze zwischen einer Bearbeitung, bei der das Muster erhalten bleibt, und einer Bearbeitung, bei der es verloren geht, liegt, lässt sich derzeit allerdings noch nicht verlässlich bestimmen. Anthropic hat die technischen Einzelheiten zur Erkennung bislang noch nicht vollständig veröffentlicht.

5 Was gilt, wenn Claude nur einen Text korrigiert?

Das Wasserzeichen kann nur in den Wörtern stecken, die Claude selbst auswählt. Bei einer reinen Rechtschreib- oder Grammatikkorrektur ändert Claude oft zu wenige Wörter, um eine zuverlässige Erkennung zu ermöglichen.

6 Können Übersetzungen ein Wasserzeichen enthalten?

Ja. Bei einer Übersetzung wählt Claude den gesamten neuen Text selbst aus. Deshalb kann auch eine von Claude erstellte Übersetzung das Wasserzeichen tragen.

Das gilt unabhängig davon, ob der Ausgangstext ursprünglich von einem Menschen oder einem anderen KI-System erstellt wurde.

7 Gilt das auch für Programmcode?

Nur eingeschränkt. Bei Programmcode gibt es häufig nur eine technisch sinnvolle oder funktionsfähige Lösung. In diesen Fällen bestehen weniger Möglichkeiten, ein statistisches Muster in die Auswahl der Tokens einzubauen.

In Bereichen mit größeren sprachlichen Gestaltungsmöglichkeiten, etwa bei Kommentaren oder erläuternden Texten innerhalb des Codes, kann dagegen ein Wasserzeichen enthalten sein.

8 Für welche KI-Systeme von Anthropic gilt die Kennzeichnung?

KI-Systeme, die am oder nach dem 2. August 2026 eingeführt werden, unterstützen bereits bei ihrer Einführung maschinenlesbare Kennzeichnungen.

Für bereits bestehende KI-Systeme arbeitet Anthropic ebenfalls daran, solche Kennzeichnungen einzuführen. Für diese KI-Systeme gilt eine Übergangsfrist bis zum 2. Dezember 2026.

9 Lässt sich erkennen, welcher Nutzer einen Text erstellt hat?

Nein. Nach Angaben von Anthropic enthält das Wasserzeichen keine Informationen zu Nutzer, Unternehmen oder Chat.

10 Kann ich selbst prüfen, ob ein Text ein Claude-Wasserzeichen enthält?

Derzeit nicht. Anthropic arbeitet nach eigener Aussage aber an Möglichkeiten, entsprechende Markierungen zu erkennen. Andere Tools von Drittanbietern zur Erkennung von KI-Texten liefern bislang keine zuverlässigen Ergebnisse.

11 Wo kann das Wasserzeichen relevant sein?

Rechtliche Kennzeichnungspflicht und technische Erkennbarkeit sind zwei unterschiedliche Fragen.

Art. 50 Abs. 2 KI-VO verpflichtet den Anbieter eines KI-Systems, sicherzustellen, dass von dem System erzeugte synthetische Audio-, Bild-, Video- oder Textinhalte in einem maschinenlesbaren Format gekennzeichnet und als künstlich erzeugt oder manipuliert erkennbar gemacht werden.

Davon zu unterscheiden ist Art. 50 Abs. 4 KI-VO. Danach müssen Betreiber von KI-Systemen KI-generierte oder -manipulierte Texte unter bestimmten Voraussetzungen kenntlich machen. Für Texte sieht die Vorschrift eine wichtige Ausnahme vor, wenn der Text einer menschlichen Überprüfung bzw. redaktionellen Kontrolle unterzogen wurde und eine natürliche oder juristische Person die redaktionelle Verantwortung für die Veröffentlichung trägt.

Eine Ausnahme von einer Kennzeichnungspflicht des Betreibers nach Art. 50 Abs. 4 KI-VO bedeutet daher nicht, dass ein Text technisch nicht mehr als KI-generiert erkennbar sein kann.

12 Was bedeutet das nun praktisch?

- Auf unsichtbare Wasserzeichen und AI-Detektoren ist momentan kein Verlass: Ein fehlendes Wasserzeichen beweist nicht, dass der Text von einem Menschen stammt. Ein Detektor-Treffer beweist nicht sicher das Gegenteil.
- Transparenzpflichten als Betreiber eines KI-Systems bestehen weiterhin und unabhängig von möglichen Kennzeichnungen durch die Anbieter von KI-Systemen.

Quellen

<https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content>

<https://www.anthropic.com/news/claude-text-watermark>

Ansprechpartner

Dr. Sebastian Horlacher, LL.M.

Grundsatzabteilung Recht

Telefon 089-551 78-131

sebastian.horlacher@vbw-bayern.de

www.vbw-bayern.de

Hinweis:

Alle Angaben dieser Publikation beziehen sich ohne jede Diskriminierungsabsicht grundsätzlich auf alle Geschlechter.