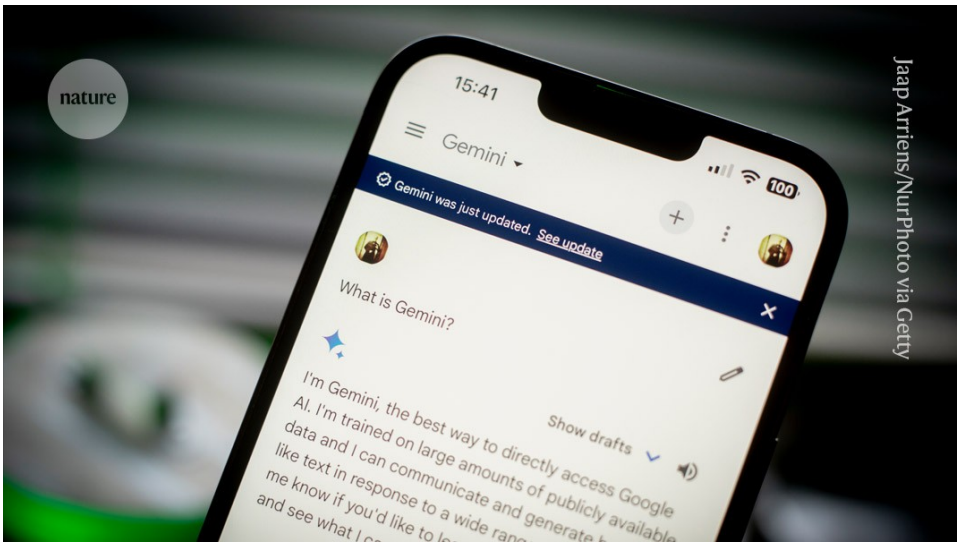
Google präsentiert unsichtbares Wasserzeichen für KI-generierte Texte

Google DeepMind hat ein unsichtbares Wasserzeichen für KI-generierte Texte entwickelt, um Falschinformationen zu bekämpfen.



Forscher bei Google DeepMind in London haben ein „Wasserzeichen“ entwickelt, um Text, der von künstlicher Intelligenz (KI) erzeugt wird, unsichtbar zu kennzeichnen – dieses wurde bereits bei Millionen von Chatbot-Nutzern eingesetzt.

Das Wasserzeichen, das am 23. Oktober in der Fachzeitschrift Nature veröffentlicht wurde[1], ist nicht das erste, das für KI-generierte Texte erstellt wurde. Es ist jedoch das erste, das in einem groß angelegten, realen Kontext demonstriert wird. „Meiner Meinung nach ist die wichtigste Neuigkeit hier, dass sie es tatsächlich einsetzen“, sagt Scott Aaronson, Informatiker an der University of Texas in Austin, der bis August an Wasserzeichen bei OpenAI gearbeitet hat, den Machern von

ChatGPT mit Sitz in San Francisco, Kalifornien.

Die Erkennung von KI-generierten Texten wird immer wichtiger, da sie eine potenzielle Lösung für die Probleme von **Fake News** und **akademischem Betrug** darstellt. Zudem könnte es dazu beitragen, **zukünftige Modelle vor der Abwertung zu schützen, indem sie nicht mit KI-generiertem Inhalt trainiert werden**.

In einer umfangreichen Studie bewerteten Nutzer des Google Gemini Large Language Model (LLM) in 20 Millionen Antworten wasserzeichenbehaftete Texte als gleichwertig mit unmarkierten Texten. „Ich bin begeistert zu sehen, dass Google diesen Schritt für die Tech-Community unternimmt“, sagt Furong Huang, Informatiker an der University of Maryland in College Park. „Es ist wahrscheinlich, dass die meisten kommerziellen Tools in naher Zukunft Wasserzeichen enthalten werden“, fügt Zakhar Shumaylov, Informatiker an der Universität Cambridge, UK, hinzu.

Wortwahl

Es ist schwieriger, ein Wasserzeichen auf Text anzuwenden als auf Bilder, da die Wortwahl im Wesentlichen die einzige Variable ist, die verändert werden kann. DeepMinds Wasserzeichen – genannt SynthID-Text – verändert, welche Wörter das Modell wählt, in einer geheimen, aber formelmäßigen Weise, die mit einem kryptografischen Schlüssel erfasst werden kann. Im Vergleich zu anderen Herangehensweisen ist DeepMinds Wasserzeichen geringfügig einfacher zu erkennen, und die Anwendung verzögert die Texterstellung nicht. „Es scheint, dass es die Konzepte der Wettbewerber beim Wasserzeichen von LLMs übertrifft“, sagt Shumaylov, der ein ehemaliger Mitarbeiter und Bruder eines der Autoren der Studie ist.

Das Tool wurde auch offengelegt, sodass Entwickler ihr eigenes Wasserzeichen auf ihre Modelle anwenden können. „Wir hoffen, dass andere Entwickler von KI-Modellen dies übernehmen und in

ihre eigenen Systeme integrieren“, sagt Pushmeet Kohli, Informatiker bei DeepMind. Google hält seinen Schlüssel geheim, damit die Nutzer keine Erkennungstools verwenden können, um wasserzeichenbehafteten Text des Gemini-Modells zu identifizieren.

Regierungen setzen **auf Wasserzeichen als Lösung zur Verbreitung von KI-generiertem Text**. Dennoch gibt es viele Probleme, darunter die Verpflichtung der Entwickler zur Verwendung von Wasserzeichen und die Koordinierung ihrer Ansätze. Anfang dieses Jahres zeigten Forscher am Eidgenössischen Technikum Zürich, dass **jedes Wasserzeichen anfällig für das Entfernen** ist, ein Prozess, der als „Scrubbing“ bezeichnet wird, oder „Spoofing“, bei dem Wasserzeichen auf Texte angewendet werden, um den falschen Eindruck zu erwecken, dass sie KI-generiert sind.

Token-Turnier

DeepMinds Ansatz basiert auf einer **bestehenden Methode**, die ein Wasserzeichen in einen Sampling-Algorithmus integriert, einen Schritt bei der Texterstellung, der vom LLM selbst getrennt ist.

Ein LLM ist ein Netzwerk von Assoziationen, das durch das Training mit Milliarden von Wörtern oder Wortteilen, die als Token bekannt sind, aufgebaut wird. Wenn ein Text eingegeben wird, weist das Modell jedem Token in seinem Vokabular eine Wahrscheinlichkeit zu, das nächste Wort im Satz zu sein. Die Aufgabe des Sampling-Algorithmus besteht darin, gemäß einer Reihe von Regeln auszuwählen, welches Token verwendet werden soll.

Der SynthID-Text-Sampling-Algorithmus verwendet einen kryptografischen Schlüssel, um jedem möglichen Token zufällige Werte zuzuweisen. Kandidatentoken werden proportional zu ihrer Wahrscheinlichkeit aus der Verteilung gezogen und in ein „Turnier“ eingeordnet. Dort vergleicht der Algorithmus die Werte

in einer Reihe von Eins-gegen-Eins-K.o.-Runden, wobei der höchste Wert gewinnt, bis nur noch ein Token übrig bleibt, das für den Text ausgewählt wird.

Diese ausgeklügelte Methode erleichtert die Erkennung des Wasserzeichens, da der gleiche kryptografische Code auf generierten Text angewendet wird, um nach den hohen Werten zu suchen, die auf „gewinnende“ Tokens hinweisen. Dies könnte auch die Entfernung erschweren.

Die mehreren Runden im Turnier können als eine Kombination aus Lock betrachtet werden, bei dem jede Runde eine andere Zahl darstellt, die gelöst werden muss, um das Wasserzeichen zu entsperren oder zu entfernen, sagt Huang. „Dieser Mechanismus macht es erheblich schwieriger, das Wasserzeichen zu scruben, zu spoofen oder zurückzuentwickeln“, fügt sie hinzu. Bei Texten mit etwa 200 Tokens zeigten die Autoren, dass sie das Wasserzeichen weiterhin erkennen konnten, selbst wenn ein zweites LLM verwendet wurde, um den Text zu umschreiben. Bei kürzeren Texten ist das Wasserzeichen weniger robust.

Die Forscher haben nicht untersucht, wie gut das Wasserzeichen gegen absichtliche Versuche zur Entfernung resistent ist. Die Widerstandsfähigkeit von Wasserzeichen gegen solche Angriffe ist eine „massive politische Frage“, sagt Yves-Alexandre de Montjoye, Informatiker am Imperial College London. „Im Kontext der KI-Sicherheit ist unklar, inwieweit dies Schutz bietet“, erklärt er.

Kohli hofft, dass das Wasserzeichen zunächst dazu beitragen wird, den gut gemeinten Einsatz von LLMs zu unterstützen. „Die guiding philosophy war, dass wir ein Werkzeug entwickeln wollen, das von der Gemeinschaft verbessert werden kann“, fügt er hinzu.

1. Dathathri, S. et al. Nature 634, 818–823 (2024).