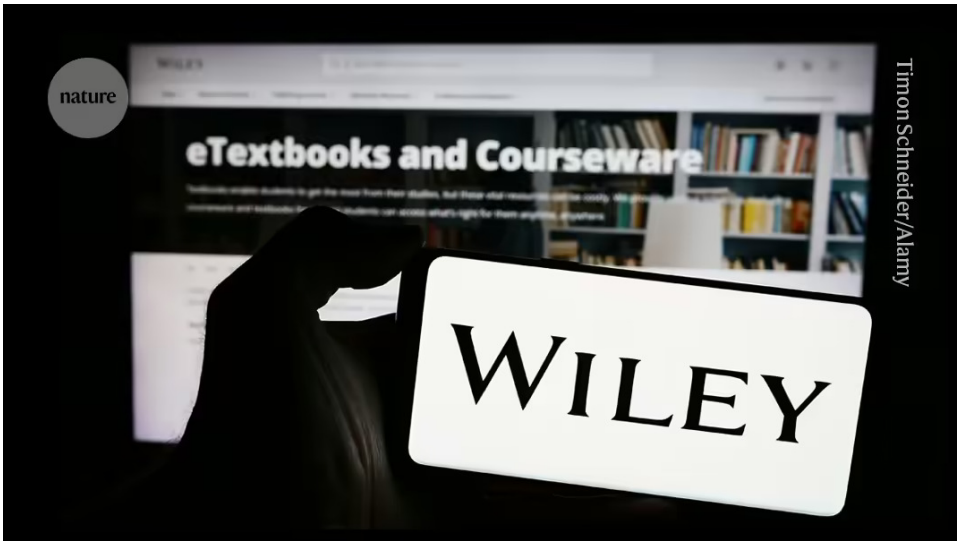
# Wurde Ihr Papier zur Schulung eines KI-Modells verwendet? Sehr wahrscheinlich

Erfahren Sie mehr über die Auswirkungen des Verkaufs von Forschungsarbeiten an Technologieunternehmen zur Schulung von KI-Modellen und die Fragen, die sich daraus ergeben. Lesen Sie, wie akademische Verlage Daten an Technologiefirmen verkaufen und welche Bedenken dies bei Forschern hervorruft.




Wissenschaftsverlage verkaufen den Zugang zu Forschungsarbeiten an Technologieunternehmen, um künstliche Intelligenz (KI)-Modelle zu trainieren. Einige Forscher haben mit Bestürzung auf solche Deals reagiert, die ohne die Konsultation der Autoren stattfinden. Der Trend wirft Fragen zur Verwendung von veröffentlichten und manchmal urheberrechtlich geschützten Arbeiten auf, um die wachsende Anzahl von in Entwicklung befindlichen KI-Chatbots zu trainieren.

Experten sagen, dass ein Forschungspapier, das noch nicht verwendet wurde, um ein großes Sprachmodell zu trainieren,

wahrscheinlich bald verwendet wird. Forscher erkunden technische Möglichkeiten für Autoren, um festzustellen, ob ihr Inhalt verwendet wird.

Letzten Monat wurde bekannt, dass der britische Wissenschaftsverlag Taylor & Francis mit Sitz in Milton Park, Großbritannien, einen zehn Millionen US-Dollar-Deal mit Microsoft unterzeichnet hat, der es dem US-Technologieunternehmen ermöglicht, auf die Daten des Verlags zuzugreifen, um seine KI-Systeme zu verbessern. Im Juni zeigte ein Investorenupdate, dass der US-Verlag Wiley 23 Millionen US-Dollar verdient hat, indem er einem nicht genannten Unternehmen erlaubte, generative KI-Modelle auf seinen Inhalten zu trainieren.

Alles, was online verfügbar ist – ob in einem Open-Access-Repository oder nicht – wurde „ziemlich wahrscheinlich“ bereits in ein großes Sprachmodell eingespeist, sagt Lucy Lu Wang, eine KI-Forscherin an der University of Washington in Seattle. „Und wenn ein Papier bereits als Trainingsdaten in einem Modell verwendet wurde, gibt es keinen Weg, dieses Papier nach dem Training des Modells zu entfernen“, fügt sie hinzu.

## Massive Datensätze

LLMs werden an riesigen Datenmengen trainiert, die häufig aus dem Internet abgeschöpft werden. Sie ermitteln Muster zwischen den oft Milliarden von Sprachausschnitten in den Trainingsdaten, sogenannten Token, die es ihnen ermöglichen, Texte mit erstaunlicher Flüssigkeit zu generieren.

Generative KI-Modelle verlassen sich darauf, Muster aus diesen Datenmassen aufzunehmen, um Texte, Bilder oder Computercode auszugeben. Wissenschaftliche Arbeiten sind für LLM-Entwickler aufgrund ihrer Länge und „hohen Informationsdichte“ wertvoll, sagt Stefan Baack, der an der Mozilla Foundation in San Francisco, Kalifornien, die Analyse von KI-Trainingsdatensätzen durchführt.

Die Tendenz zum Kauf hochwertiger Datensätze wächst. Dieses Jahr hat die *Financial Times* ihr Material dem **ChatGPT-Entwickler OpenAI** in einem lukrativen Deal angeboten, ebenso wie das Online-Forum Reddit an Google. Und da wissenschaftliche Verlage die Alternative wahrscheinlich als unerlaubtes Abschöpfen ihrer Arbeit betrachten, „denke ich, dass noch mehr solcher Deals bevorstehen“, sagt Wang.

## Geheimnisse der Information

Einige KI-Entwickler, wie das Large-scale Artificial Intelligence Network, halten ihre Datensätze absichtlich offen, aber viele Unternehmen, die generative KI-Modelle entwickeln, haben einen Großteil ihrer Trainingsdaten geheim gehalten, sagt Baack. „Wir haben keine Ahnung, was darin ist“, sagt er. Open-Source-Repositories wie arXiv und die wissenschaftliche Datenbank PubMed gelten als „sehr beliebte“ Quellen, obwohl paywalled Journalartikel wahrscheinlich von großen Technologieunternehmen kostenlos zu lesenden Abstracts abgeschöpft werden. „Sie sind immer auf der Jagd nach solchen Informationen“, fügt er hinzu.

Es ist schwierig nachzuweisen, dass ein LLM ein bestimmtes Papier verwendet hat, sagt Yves-Alexandre de Montjoye, ein Informatiker am Imperial College London. Eine Möglichkeit besteht darin, das Modell mit einem ungewöhnlichen Satz aus einem Text zu konfrontieren und zu prüfen, ob die Ausgabe mit den nächsten Worten im Original übereinstimmt. Wenn dies der Fall ist, ist das ein gutes Zeichen dafür, dass das Papier im Trainingsset enthalten ist. Wenn nicht, bedeutet das nicht, dass das Papier nicht verwendet wurde – nicht zuletzt, weil Entwickler das LLM programmieren können, um die Antworten zu filtern, um sicherzustellen, dass sie nicht zu eng mit den Trainingsdaten übereinstimmen. „Es braucht viel, damit das funktioniert“, sagt er.

Ein weiteres Verfahren zur Überprüfung, ob Daten in einem Trainingsdatensatz enthalten sind, wird als

Mitgliedschaftsinferezangriff bezeichnet. Dies basiert auf der Idee, dass ein Modell selbstsicherer über seine Ausgabe ist, wenn es etwas sieht, was es zuvor gesehen hat. De Montjoyes Team hat eine Version davon entwickelt, die als Copyright-Falle bezeichnet wird, für LLMs.

Um die Falle zu stellen, generiert das Team plausible, aber unsinnige Sätze und versteckt sie in einem Werk, beispielsweise als weißen Text auf weißem Hintergrund oder in einem Feld, das auf einer Webseite als Nullbreite angezeigt wird. Wenn ein LLM von einem unbenutzten Kontrollsatz „überrascht“ ist – ein Maß für seine Verwirrung -, mehr als von dem im Text versteckten Satz, „ist das statistischer Nachweis, dass die Fallen zuvor gesehen wurden“, sagt er.

## Urheberrechtliche Fragen

Auch wenn es möglich wäre nachzuweisen, dass ein LLM auf einem bestimmten Text trainiert wurde, ist nicht klar, was als nächstes passiert. Verlage behaupten, dass die Verwendung urheberrechtlich geschützter Texte im Training ohne Lizenzierung als Verletzung gilt. Aber ein rechtliches Gegenargument besagt, dass LLMs nichts kopieren – sie extrahieren Informationsgehalt aus den Trainingsdaten, der zerkleinert wird, und nutzen ihr gelerntes Wissen, um neuen Text zu generieren.

Möglicherweise könnte ein Gerichtsverfahren dazu beitragen, dies zu klären. In einem laufenden US-Urheberrechtsfall, der wegweisend sein könnte, verklagt *The New York Times* Microsoft und den Entwickler von ChatGPT, OpenAI, in San Francisco, Kalifornien. Die Zeitung wirft den Unternehmen vor, ihren journalistischen Inhalt ohne Erlaubnis zur Schulung ihrer Modelle verwendet zu haben.

Viele Akademiker sind glücklich, wenn ihre Arbeit in die Trainingsdaten von LLMs aufgenommen wird – insbesondere, wenn die Modelle genauer werden. „Ich persönlich habe nichts

dagegen, wenn ein Chatbot in meinem Stil schreibt“, sagt Baack. Aber er räumt ein, dass sein Beruf nicht von den Ausgaben der LLMs bedroht ist, wie die anderer Berufe, wie Künstler und Schriftsteller, es sind.

Individuelle wissenschaftliche Autoren haben derzeit nur wenig Einfluss, wenn der Verlag ihres Papiers den Zugang zu ihren urheberrechtlich geschützten Werken verkauft. Für öffentlich abrufbare Artikel gibt es keine etablierten Mittel, um eine Gutschrift zuzuteilen oder zu wissen, ob ein Text verwendet wurde.

Einige Forscher, darunter de Montjoye, sind frustriert. „Wir wollen LLMs, aber wir wollen immer noch etwas, das fair ist, und ich glaube, wir haben noch nicht erfunden, wie das aussieht“, sagt er.

Details