ZEIT ONLINE

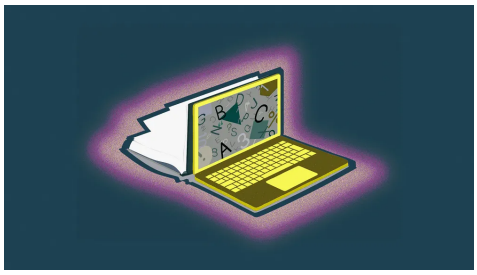
Künstliche Intelligenz

Das gestohlene Wort

Knapp 190.000 Bücher aus illegalen Quellen kamen für das Training zahlreicher KI-Sprachmodelle zum Einsatz. Viele Autoren sind empört, können aber nur wenig ausrichten.

Von **Eike Kühl**

30. September 2023, 15:08 Uhr / 57 Kommentare /



197.000 Bücher, von Shakespeare zu Stephen King, stecken in Books3. Und damit auch in vielen KI-Modellen. © [M] Alice Mollon/ikon images/ imago images

Große Sprachmodelle wie GPT-3 von OpenAI, Metas LLaMA und Googles Bard sind Datenfresser. Damit sie später grammatikalisch richtige und inhaltlich schlüssige Sätze bilden können, müssen sie mit sogenannten Trainingsdaten gefüttert werden. Die sind zu großen Teilen aus dem öffentlichen Internet zusammengeklaubt [https://www.zeit.de/digital/datenschutz/2023-08/ki-datenschutz-trainingsdaten-zugriff-chatgpt], aber nicht nur: Manche Datensätze, mit denen einige bekannte Sprachmodelle trainiert wurden, bestehen nachweislich aus Raubkopien.

Im Mittelpunkt der aktuellen Debatte steht ein Datensatz namens Books3. Darin enthalten sind 197.000 Ausgaben größtenteils westlicher Literatur, von Shakespeare zu Stephen King, von Kochbüchern zur Philosophie von Sartre, und von Nobelpreisträger Günter Grass zu US-Comedian Sarah Silverman. Jene Sarah Silverman, die wie mittlerweile zahlreiche weitere Autoren [https://www.zeit.de/digital/2023-09/openai-kuenstliche-intelligenz-klage-schriftsteller], Unternehmen wie Meta und OpenAI beschuldigt, ihre urheberrechtlich geschützten Werke ohne Erlaubnis zu nutzen.

Das US-Magazin *The Atlantic* hat am Montag eine durchsuchbare Datenbank veröffentlicht [https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/], die rund 183.000 der in Books3 enthaltenen Titel enthält. Die Journalisten haben die rund 100 Gigabyte große Datenbank, bei der es sich um eine gigantische Textdatei handelt, auf ISBN hin untersucht. So konnten sie herausfinden, welche Werke genau enthalten sind. Man darf davon ausgehen, dass jede künstliche Intelligenz, die Books3 in ihren Trainingsdaten verwendet, mit diesen Büchern gefüttert wurde.

Eines dieser Sprachmodelle ist LLaMA des Facebook-Mutterkonzerns Meta. Das geht aus der offiziellen Dokumentation hervor [https://arxiv.org/pdf/2302.13971.pdf]. Auch das Sprachmodell GPT-J [https://huggingface.co/EleutherAI/gpt-j-6b] der gemeinnützigen Forschungsgruppe EleutherAI sowie BloombergGPT [https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/] des gleichnamigen Medienunternehmens benutzen den Datensatz. Auf welchen Trainingsdaten GPT-4 basiert, ist dagegen unklar; OpenAI schweigt [https://cdn.openai.com/papers/gpt-4.pdf] "aufgrund des Wettbewerbs und den Auswirkungen auf die Sicherheit" dazu. Im Fall von GPT-3, das aktuell die Basis der kostenlosen Version des Chatbots ChatGPT bildet, waren aber die Datensätze Books1 und Books2 enthalten [https://lambdalabs.com/blog/demystifying-gpt-3], die sich mutmaßlich aus sogenannten Schattenbibliotheken, also aus nicht öffentlichen und häufig auch nicht legalen Sammlungen speisen.

Die Hintergründe von Books3

Aus einer solchen Schattenbibliothek stammt auch Books3. Im Oktober 2020, wenige Wochen nachdem OpenAI erstmals GPT-3 der Öffentlichkeit präsentiert hatte, begann der KI-Forscher Shawn Presser damit, den Datensatz zusammenzutragen. Wie er damals auf X, vormals Twitter, schrieb [https://twitter.com/theshawwn/status/1320282149329784833] und in aktuellen Interviews beteuert [https://www.wired.com/story/battle-over-books3/], sei seine Absicht lediglich gewesen, KI-Projekte, die weniger Kapital als OpenAI haben, mit guten Trainingsdaten zu versorgen, um konkurrenzfähig zu bleiben.

Die fast 200.000 Bücher suchte Presser nicht selbst zusammen, sondern bediente sich bei Bibliotik, einem Bittorrent-Tracker (wie die deutlich bekanntere Pirate Bay), der auf E-Books spezialisiert ist. Dort teilen die Nutzerinnen und Nutzer untereinander Werke, die in den meisten Fällen urheberrechtlich geschützt sind. Sie verbreiten also Raubkopien. Pressers Vorgehen kam auch deshalb in der Trackerszene nicht besonders gut an [https://www.reddit.com/r/trackers/comments/12numrs/til_bibliotik_used_to_train_llm_ais/], da er Bibliotik namentlich als Quelle für seinen Datensatz erwähnte.

Im Dezember 2020 wurde Books3 von der Forschungsgruppe EleutherAI i [https://arxiv.org/pdf/2101.00027.pdf]n einen noch viel größeren Trainingsdatensatz namens *The Pile* aufgenommen [https://arxiv.org/pdf/2101.00027.pdf]. Auch wenn Books3 letztlich

nur einen kleinen Teil von *The Pile* und der Trainingsdaten insgesamt ausmacht (im Fall von LLaMA sind es nur 4,5 Prozent [https://arxiv.org/pdf/2302.13971.pdf]), lässt sich nicht abstreiten, dass einige der führenden KI-Sprachmodelle Daten aus illegalen Quellen, oder salopp gesagt gestohlene Wörter, enthalten.

Ungeklärte Rechtsfragen bei KI-Sprachmodellen

Für klagende Autorinnen wie Sarah Silverman und deren Anwälte ist die Sache deshalb klar. Wie sie schreiben [https://www.documentcloud.org/documents/23869675-kadrey-meta-complaint?responsive=1&title=1], verstoßen Unternehmen wie Meta und OpenAI gegen das Urheberrecht, indem sie die Daten aus unrechtmäßigen Quellen und ohne Einwilligung der Urheberinnen und Urheber zum Training ihrer Algorithmen verwenden.

Tatsächlich ist die Sache komplexer. So könnten sich die Unternehmen auf die in den USA geltende *Fair-Use*-Regel beziehen: Sprachmodelle kopieren nicht den Inhalt der Trainingsdaten, sondern schaffen stattdessen neue Inhalte in Form von Antworten auf Nutzerfragen. Da diese Werke nicht in direkter Konkurrenz zu den Originalbüchern stehen, können auch keine urheberrechtlichen Ansprüche geltend gemacht werden.

Fraglich ist außerdem, ob es überhaupt eine Rolle spielt, ob die Daten von Books3 aus einer illegalen Schattenbibliothek stammen. "Wenn die Quelle nicht autorisiert ist, kann das ein Faktor sein", sagt der der US-Rechtsprofessor Jason Schultz gegenüber *The Atlantic*. Wie das Magazin *Wired* schreibt [https://www.wired.com/story/artificial-intelligence-copyright-law/], gebe es aber in den USA "keinen Präzedenzfall, der *Fair Use* direkt davon abhängig macht, ob die urheberrechtlich geschützten Werke legal erworben wurden oder nicht." Einmal mehr zeigt sich, dass zahlreiche Urheberrechtsfragen im Kontext künstlicher Intelligenz [https://www.zeit.de/digital/internet/2023-01/ki-bildgeneratoren-klage-kuenstler-urheberrecht-verletzung] noch nicht abschließend beantwortet wurden.

Die KI-Urheberrechtsdebatte breitet sich aus

Ungeachtet dessen legt der Fall von Books3 auch die Machtverhältnisse in der KI-Branche offen. Laut Shawn Presser könnten sich nur große Unternehmen wie OpenAI solche Datensätze leisten, wenn sie dafür bezahlen oder Lizenzen einholen müssten. Kleinere Unternehmen und Start-ups hätten diese Möglichkeiten nicht. Das könnte dazu führen, dass sich die Marktmacht konzentriert, sagte er im Gespräch mit *Wired* [https://www.wired.com/story/battle-over-books3/]. Dazu kommt: Wer bereits seine Algorithmen mit Books3 gefüttert hat, kann die Daten nicht nachträglich herausnehmen und hat deshalb womöglich einen Vorteil gegenüber neuen Sprachmodellen, die auf Books3 und andere urheberrechtlich bedenkliche Datensätze verzichten.

Der Sog der Debatte breitet sich derweil aus. Vergangene Woche hat Shawn Presser Post von Silvermans Anwälten erhalten, wie er auf X schreibt [https://twitter.com/theshawwn/status/1704559992135717238]. Sie weisen ihn darauf hin, dass er E-Mails und Dokumente sichern sollte, die für den Fall relevant sein könnten. Die dänische Anti-Piraterie-Organisation Rights Alliance geht derzeit verstärkt gegen Plattformen vor [https://torrentfreak.com/anti-piracy-group-takes-prominent-ai-training-dataset-books3-offline-230816/], die den Books3-Datensatz zum Download anbieten. Und die US-amerikanische Authors Guild hat in einer Petition bereits mehr als 10.000 Unterschriften von Autorinnen und Autoren gesammelt, in der sie die KI-Firmen auffordern, das Einverständnis von Autoren einzuholen und diese fair zu bezahlen, wenn sie ihre Werke verwenden.

Manche Autoren beugen sich allerdings der Entwicklung. Das bekannteste Beispiel ist Stephen King, dessen Worte nach denen von Shakespeare am häufigsten in Books3 enthalten sind. Kreative künstliche Intelligenz übe auf ihn eine "schreckliche Faszination" aus, schreibt King [https://www.theatlantic.com/books/archive/2023/08/stephen-king-books-ai-writing/675088/]. Ob er deshalb, wenn er könnte, verbieten würde, dass eine Maschine aus seinen Worten lernt? "Dann könnte ich auch ein Luddit sein, der versucht, den industriellen Fortschritt zu stoppen, indem er eine Webmaschine zertrümmert."