KI und Urheberrecht – was ist beim Einsatz von LLMs erlaubt?

Vor einigen Monaten haben neue Player die Rechtsszene betreten und seitdem für zahlreiche juristische Diskussionen gesorgt: Large Language Models (LLM), besser bekannt als ChatGPT, Azure OpenAI, PaLM 2 und Co. Insbesondere im Urheberrecht wirft die generative KI viele Fragen auf: Wann verletzen mit LLMs generierte Texte Urheberrechte? Und können an solchen Texten eigene Urheberrechte entstehen? Erlaubt das Urheberrecht die Vervielfältigung und Speicherung von Daten zum Training von LLMs?

LLMs generieren Texte. Sie wurden mit großen Datenmengen trainiert. Sie erstellen Texte, indem sie mithilfe ihrer Trainingsdaten jeweils die nächsten Elemente des generierten Textes selbst vorhersagen. Das LLM berechnet die Wahrscheinlichkeit von Wortfolgen – beziehungsweise von Folgen von Tokens – und entwickelt diese im nächsten Schritt selbständig zu Texten weiter. Die Antworten, also der Output der LLMs, basieren auf der wahrscheinlichsten Wortfolge, die aus den Wörtern der Eingabe (Prompt) berechnet wird.

Kann der Output Urheberrechte verletzen?

Eine gute Nachricht: LLMs sind nicht darauf ausgelegt, Plagiate zu erstellen. Im Gegensatz zu Internet-Suchmaschinen suchen sie nicht nach bereits vorhandenen Texten und zeigen diese an, sondern generieren neue Texte. Je nach den Anweisungen, die Nutzer:innen dem LLM erteilen, kann das allerdings trotzdem Urheberrechte verletzen. Es besteht nämlich die Gefahr, dass die KI Output generiert, der mit einem urheberrechtlich geschützten Werk identisch ist, mit dem sie trainiert wurde.

In folgenden Beispielen kann die generative KI Urheberrechte verletzen:

- Wenn die KI gebeten wird, einen bestimmten Text zu reproduzieren, zum Beispiel einen Liedtext, der noch nicht gemeinfrei ist,
- wenn die KI urheberrechtlich geschützte Texte lediglich in eine andere Sprache übersetzt und
- wenn die KI Teile eines Textes reproduziert, die einen eigenen urheberrechtlichen Schutz genießen.

Da technische, informative Texte meistens nicht urheberrechtlich geschützt sind, ist eine Verletzung des Urheberrechts bei der Verarbeitung solcher Texte weniger wahrscheinlich. Theoretisch möglich bleibt sie aber dennoch: Ist es dem Autor oder der Autorin des technischen Textes gelungen, bei der Abfassung Kreativität zu zeigen, wäre unter Umständen auch an einem Sachtext ein Urheberrecht entstanden. Und wenn die KI gerade diese Elemente übernimmt, wäre es verletzt.

Können an KI-generierten Texten Urheberrechte entstehen?

Die nächste Frage, die sich im Zusammenhang mit GenAI und Urheberrecht stellt, ist die nach dem urheberechtlichen Schutz des generierten Outputs: Kann mithilfe von KI ein Werk entstehen, das urheberrechtlichen Schutz genießt? Und wenn ja, wem steht das Urheberrecht dann zu? Auch hier gibt es keine allgemeingültige Antwort.

Nach deutschem Recht sind nur persönliche geistige Schöpfungen schutzfähig. Die Schöpfung muss aus dem

Gedankengang eines Menschen resultieren und das Ergebnis eines zielgerichteten geistigen Schaffensprozesses sein. Zufällige Ergebnisse, wie zum Beispiel unbeabsichtigte Farbspritzer oder ein von einem Affen geknipstes Foto, können keinen urheberrechtlichen Schutz beanspruchen. Nach deutschem Recht kann nur ein Mensch mit menschlicher Intelligenz und nicht eine KI als Urheberin oder Urheber gelten, und nur der Mensch kann urheberrechtlich geschützte Werke schaffen. Entscheidend ist, dass Urheber:innen in ihren kreativen Entscheidungen frei sind.

Bei der Verwendung von LLMs, wie wir sie bereits gewohnt sind, findet in der Regel eher keine hinreichend kreative Entscheidung der menschlichen Nutzer:innen der KI statt. Der geschriebene Prompt kann ein urheberrechtlich geschütztes Werk sein, führt aber nicht zum Schutz des durch die KI generierten Outputs. Nutzer:innen haben in der Regel keinen wesentlichen Einfluss auf die maschinelle Ausführung, die eigentliche Produktion des Textes.

Es kann jedoch Fälle geben, in denen eine andere Beurteilung gerechtfertigt ist, nämlich dann, wenn die Nutzer:innen das LLM als ein Werkzeug einsetzen und betreiben, das lediglich ihre persönliche kreative Absicht umsetzt. Etwas anschaulicher wäre das vergleichbar mit der Verwendung eines Pinsels. Wenn der Pinsel lediglich über das Papier rollt, zum Beispiel weil er fallen gelassen wird, entsteht kein urheberrechtlich geschütztes Werk, auch wenn dabei Farbe auf dem Papier zurückbleibt. Schwingt ein Maler den Pinsel jedoch absichtlich in einer bestimmten Weise, kann ein geschütztes Gemälde entstehen. Wenn KI in einer vergleichbaren Weise eingesetzt wird, kann tatsächlich ein urheberrechtlich geschütztes Werk entstehen.

Das wirft sofort die Frage auf, wer Urheber:in dieses Werkes ist und wem die Rechte daran zustehen. Hier sind verschiedene Lösungen möglich. Es könnte allein der Nutzer beziehungsweise die Nutzerin der KI sein, oder es könnte ein gemeinsames Werk zwischen dem Nutzer/der Nutzerin und dem KI-Programmierer sein. Diese Frage wird die Urheberrechtsexpert:innen in den verschiedenen Rechtsordnungen sicher noch einige Zeit beschäftigen.

Dürfen Daten zu Trainingszwecken von KI vervielfältigt und gespeichert werden?

Eine weitere zentrale Frage: Erlaubt das Urheberrecht die Vervielfältigung und Speicherung von Daten zum Training von KI-Systemen? Und falls ja, wie lange dürften diese Daten dann aufbewahrt werden?

Zulässig ist seit 2021 die Vervielfältigung von rechtmäßig zugänglichen Werken zum Zwecke des Text- und Data-Minings gem. § 44b Abs. 2 UrhG. Das bedeutet, dass digitale oder digitalisierte Werke automatisiert analysiert werden, um Informationen, insbesondere Muster, Trends und Korrelationen, zu extrahieren. Dabei werden die Daten gespeichert, also vervielfältigt. Die Daten müssen jedoch gelöscht werden, wenn sie nicht mehr für Text- und Data-Mining benötigt werden. Aber gilt das auch für das Training einer KI? Zu dieser Frage gibt es noch keine Rechtsprechung. Die Gesetzesbegründung deutet zwar darauf hin, dass § 44b UrhG die Vervielfältigung und Speicherung von rechtmäßig zugänglichen Daten für das KI-Training grundsätzlich erlaubt. Auch wenn der Gesetzgeber damals vermutlich noch nicht an Large Language Models gedacht hat, sind diese unserer Meinung nach erfasst. Denn die Vorschrift berücksichtigt auch die Interessen der Urheber:innen angemessen, da sie sich eine solche Nutzung vorbehalten und sie auch verbieten können.

Es stellt sich jedoch noch eine andere Frage: Wie lange dürfen solche Trainingsdaten gespeichert werden? Gibt es

eine zeitliche Begrenzung, nach der die Daten gelöscht werden müssen, oder besteht der rechtfertigende Zweck so lange fort, wie die KI in Betrieb ist? Auf diese Frage gibt es noch keine endgültige Antwort. Es bleibt abzuwarten, wie sich Gesetzgebung und Rechtsprechung in diesen Bereichen in den kommenden Jahren entwickeln werden.

Fazit

Die Nutzung von LLMs wirft rechtliche Fragen hinsichtlich der möglichen Verletzung von Urheberrechten, der Entstehung neuer Urheberrechte und der Zulässigkeit der Vervielfältigung und Speicherung von Daten für das Training von LLMs auf.

Das Training der KI mittels rechtmäßig zugänglicher Daten ist unseres Erachtens erlaubt, da die Interessen der Urheberrechtsinhaber angemessen berücksichtigt werden. Ungeklärt ist allerdings, wie lange diese Daten gespeichert werden dürfen.

Wie sich die Rechtsprechung zu diesen Fragen positionieren wird, und ob der Gesetzgeber hier nochmal tätig werden wird, bleibt abzuwarten.

Ansprechpartner:

Francois Heynike, LL.M. (Stellenbosch)
Tel: +49-69-951195770
fheynike@kpmg-law.com

Dr. Daniel Taraz
Tel: +49 40 360994-5483
danieltaraz@kpmg-law.com