

Dokument	sic! 2025 S. 381
Autor	Florent Thouvenin
Titel	Retrieval-Augmented Generation (RAG) – ein urheberrechtliches 101
Seiten	381-392
Publikation	Zeitschrift für Immaterialgüter-, Informations- und Wettbewerbsrecht
Herausgeber	Mathis Berger, Jürg Borer, Pranvera Këllezi, Anne-Virginie La Spada, Roland Mathys, Peter Georg Picht, Cyrill P. Rigamonti, Florent Thouvenin, Jacques de Werra, Gregor Wild
Frühere Herausgeber	Marc Amstutz, Ivan Cherpillod, Reto M. Hilty, Michel Jaccard, Eugen Marbach, Jürg Müller, Michael Ritscher, Werner Stieger, Rolf H. Weber
ISSN	1422-2019
Verlag	Helbing Lichtenhahn Verlag

Florent Thouvenin

Retrieval-Augmented Generation ([RAG](#)) – ein urheberrechtliches 101

Retrieval-Augmented Generation ([RAG](#)) wird verwendet, um die Zuverlässigkeit und Aktualität des Outputs von Sprachmodellen zu erhöhen. [RAG](#) ist weit verbreitet, wurde aber noch kaum urheberrechtlich untersucht. Dieser Beitrag vermittelt einen Überblick über die Rechtsfragen und nimmt eine erste Einordnung vor.&cbr;

La génération augmentée de récupération (RAG, Retrieval-Augmented Generation) est utilisée pour augmenter la fiabilité et l'actualité des réponses fournies par les modèles de langage. La RAG est très répandue, mais elle n'a guère été étudiée sous l'angle du droit d'auteur. Cet article donne un aperçu des questions juridiques et procède à une première classification.

Florent Thouvenin, Prof. Dr. iur., Rechtsanwalt, Ordinarius für Informations- und Kommunikationsrecht und Vorsitzender des Leitungsausschusses des Center for Information Technology, Society, and Law (ITSL) an der Universität Zürich.

Beim Verfassen der technischen Ausführungen zu [RAG](#) (Kapitel II.) wurde ChatGPT genutzt. Der Autor dankt Markus Danhel, Head AI Software – Austria & Switzerland, IBM Technology, für die Klärung technischer Fragen und seinen wissenschaftlichen Mitarbeitenden, Viviane Ammann, MLaw, LL.M., und Lennart Deutschmann, M.iur., für die wertvolle Unterstützung bei der Recherche und beim Erstellen der Fussnoten.

I. Einleitung

Die jüngeren Entwicklungen im Bereich der Künstlichen Intelligenz (KI) stellen das Recht vor zahlreiche und vielfältige Herausforderungen. Im Vordergrund stehen dabei Fragen nach der Transparenz und Sicherheit von KI-Systemen, die Gefahr von Diskriminierung und Manipulation beim Einsatz solcher Systeme und

Herausforderungen im Bereich des Datenschutzes und der Haftung.¹ Auch das Immaterialgüterrecht ist herausgefordert. Einige Fragen wurden schon früh und eingehend diskutiert, etwa ob Patente für Erfindungen erteilt werden können, die von KI geschaffen werden,² und ob von KI-generierte Inhalte urheberrechtlich geschützt sind.³ Bei der zweiten Frage scheint sich in den meis-

sic! 2025 S. 381, 382

ten Rechtsordnungen die Auffassung durchzusetzen, dass autonom von KI generierte Inhalte nicht urheberrechtlich geschützt sind.⁴ Eine parallele Entwicklung zeichnet sich im Patentrecht ab, in dem ebenfalls weitgehend angenommen wird, dass KI-Systeme nicht Erfinder im rechtlichen Sinn sein können.⁵

Seit dem Markteintritt von ChatGPT Ende November 2022 und dem rasch wachsenden Angebot an Systemen der sog. generativen KI steht die Frage im Vordergrund, ob urheberrechtlich geschützte Werke auch ohne Zustimmung der Rechteinhaber für das Training von (generativen) KI-Modellen verwendet werden

- 1 Siehe dazu für die Schweiz: Auslegeordnung zur Regulierung von künstlicher Intelligenz, Bericht an den Bundesrat, 12. Februar 2025; N. Braun Binder/T. Burri/M. F. Lohmann/M. Simmler/F. Thouvenin/K. N. Vokinger, Künstliche Intelligenz: Handlungsbedarf im Schweizer Recht, Jusletter vom 28. Juni 2021; F. Thouvenin/S. Volz, White Paper Transparenz, Juli 2024, <www.itsl.uzh.ch/dam/jcr:0903cf28-af7b-47e7-90b6-1eb0457c2f66/White%20Paper_Transparenz.pdf> (20. Mai 2025); zur Diskriminierung: F. Thouvenin/S. Volz, White Paper Diskriminierung, Juni 2024, <www.itsl.uzh.ch/dam/jcr:19d5ad6c-2671-431c-86d7-822e27877606/2024_ITSL_WhitePaper_Diskriminierung.pdf> (20. Mai 2025); F. Stöcklin, Robot Recruiting – Die Haftung des Arbeitgebers für Persönlichkeitsverletzungen beim Einsatz von Künstlicher Intelligenz im Bewerbungsprozess, [sui generis 2023, 1 ff.](#); zur Manipulation: F. Thouvenin/S. Volz, White Paper Manipulation, Juni 2024, <www.itsl.uzh.ch/dam/jcr:c301840c-f825-42eb-9f83-c837dabcb75a/2024_ITSL_WhitePaper_Manipulation.pdf> (20. Mai 2025); zum Datenschutz: F. Thouvenin/S. Volz, White Paper Datenschutz, Juni 2024, <www.itsl.uzh.ch/dam/jcr:537f9483-f387-4046-b104-226fe3460280/2024_ITSL_UniBas_WhitePaper_Datenschutz.pdf> (22. April 2025); A. Epiney/S. Rovelli, Künstliche Intelligenz und Datenschutz/L'intelligence artificielle et protection des données, Zürich 2021; C. Lutz/B. Domenig/A. Flükiger, Datenschutzkonformer Einsatz von ChatGPT an Schulen, [sic! 2024, 94 ff.](#); zur Haftung: C. Gordon/T. Lutz, Haftung für automatisierte Entscheidungen – Herausforderungen in der Praxis, [SZW 2020, 53 ff.](#); I. Wildhaber, Eine Einführung in die ausservertragliche Haftung für Künstliche Intelligenz (KI), in: Haftpflichtprozess 2021, 15 ff.; P. Pichonnaz, La responsabilité pour l'intelligence artificielle: un régime de responsabilité objective simple comme solution nuancée, in: E. Belser/P. Pichonnaz/H. Stöckli (éd.), Le droit sans frontières – Recht ohne Grenzen – Law without borders, Mélanges pour Franz Werro, Berne 2022, 513 ss.
- 2 Siehe dazu für die Schweiz: A. Früh/D. Haux, Künstliche Intelligenz als Erfinder:in?, Jusletter 15. November 2021; W. Straub, Immaterialgüterrechtlicher Schutz mit KI geschaffener Werke und Erfindungen, Jusletter 7. August 2023; V. Vallone, Wer hat's erfunden – Künstliche Intelligenz oder Mensch?, [sic! 2023, 203 ff.](#); F. Thouvenin/P. G. Picht, AI & IP: Empfehlungen für Rechtsetzung, Rechtsanwendung und Forschung zu den Herausforderungen an den Schnittstellen von Artificial Intelligence (AI) und Intellectual Property (IP), [sic! 2023, 507 ff.](#), 509 f.
- 3 Siehe dazu für die Schweiz: T. Rohner, Der Schutz von KI-Schöpfungen im schweizerischen Urheberrecht, ZGE 2019, 33 ff.; Thouvenin/Picht (Fn. 2), 510 f.; C. Semmelmann, Künstliche Intelligenz und Urheberrecht: Stand zu Training und Output 2024, [sic! 2024, 637 ff.](#); I. Cherpillod, Intelligence artificielle et droit d'auteur, [sic! 2023, 445 ff.](#)
- 4 Siehe dazu etwa: für die USA: U.S. Court of Appeals, D.C. Cir. vom 11. Februar 2025, 130 F.4th 1039, «Thaler v. Perlmutter»; U.S. Copyright Office, Copyright and Artificial Intelligence – Part 2: Copyrightability – A Report Of The Register Of Copyrights, <www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf> (Januar 2025); für das Vereinigte Königreich: UKIPO, Copyright and Artificial Intelligence Consultation, <www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence#d-ai-outputs> (Dezember 2024); für Deutschland: N. Luthhöft/C. Hendel, Rechte von Urhebern und Leistungsschutzberechtigten an mittels künstlicher Intelligenz geschaffener Musik, GRUR 2024, 1841 ff., 1842; M. Muhr, KI-Schöpfungen und Urheberrecht, Berlin 2022, 50 ff.; A. Lauber-Rönsberg, Autonome «Schöpfung» – Urheberschaft und Schutzfähigkeit, GRUR 2019, 244 ff., 245, m.w.H.; für die Schweiz: Cherpillod (Fn. 3), 445 ff., 449 f.; C. R. Dörfler, Das Schöpferprinzip im Immaterialgüterrecht, Bern 2024, 218 f.; P. Gilliéron, Intelligence artificielle: la titularité des données, in: Aspects juridiques de l'intelligence artificielle, Bern 2024, 13 ff., 35 f.
- 5 Für die USA z.B.: Fed. Cir. vom 5. August 2022, 43 F.4th 1207, «Thaler v. Vidal», Certiorari abgelehnt: USSC vom 24. April 2023, 143 S. Ct. 1783; für das Vereinigte Königreich: UKSC vom 20. Dezember 2023, [2023] UKSC 49, «Thaler v. Comptroller-General of Patents, Designs and Trademarks»; für Deutschland: BGH vom 11. Juni 2024, X [ZB 5/22](#), «DABUS/FOOD CONTAINER»; siehe auch die Entscheide des EPA vom 27. Januar 2020, EP 18 275 163 und EP 18 275 174, bei denen die Beschwerden in den Entscheiden der Jur. Beschwerdekammer vom 21. Dezember 2021, J 0008/20 «Designation of inventor/DABUS» und J 0009/20 «Designation of inventor/DABUS II» abgelehnt wurden; für die Schweiz: Thouvenin/Picht (Fn. 2), 507 ff., 509 ff.; Vallone (Fn. 2), 203 ff.; Straub (Fn. 2), N 16, 24 f.; Früh/Haux (Fn. 2), N 4 f.

dürfen. Diese Frage wurde und wird in zahlreichen Publikationen im In- und Ausland untersucht⁶ und mittlerweile liegen auch erste Gerichtsentscheide vor. Mit Spannung werden vor allem die Entscheide der US-amerikanischen Gerichte erwartet,⁷ die unter anderem klären werden, ob die Verwendung von urheberrechtlich geschützten Werken für das Training von KI als *fair use* zu qualifizieren ist. In einem ersten Entscheid des *District Court of Delaware* wurde dies verneint.⁸ In der EU liegt, soweit ersichtlich, erst ein Entscheid vor. In diesem ist das [LG Hamburg](#)⁹ zum Schluss gekommen, dass die Nutzung von Werken für das Training von KI von den *Text-und-Data-Mining*-Schranken erfasst wird, welche die Mitgliedstaaten nach den Vorgaben von Art. 3 und Art. 4 der DSM-Richtlinie¹⁰ in ihren Urheberrechtsgesetzen umgesetzt haben.¹¹ In der Schweiz wird derzeit vor allem diskutiert, ob diese Nutzungen von der Wissenschaftsschranke ([Art. 24d URG](#)) freigestellt wird.¹²

Diese Fragen können hier allerdings offenbleiben. Denn dieser Beitrag untersucht eine Form der Nutzung von Werken durch KI-Systeme, die bisher noch kaum rechtlich analysiert,¹³ aber bereits in den politischen Diskurs eingebracht worden ist.¹⁴ Hintergrund ist ein zentrales Problem von KI-Systemen, insb. von Chatbots: Diese neigen zu Halluzinationen, also dazu, falsche oder frei erfundene Aus-

sic! 2025 S. 381, 383

sagen zu produzieren, die zwar plausibel klingen, aber keinen Bezug zur Realität haben. Dies geschieht, weil grosse Sprachmodelle (*Large Language Models, LLMs*) Muster aus Trainingsdaten extrapolieren und auf diese Weise statistische Zusammenhänge zwischen Worten und Wortteilen (*token*) «erlernen», auf deren Grundlage sie dann voraussagen, welcher *token* angesichts der Reihe der vorangehenden *token* der wahrscheinlichste nächste *token* ist. Da Sprachmodelle auf statistischen Zusammenhängen und

⁶ Siehe dazu bspw. für die Schweiz: S. Marmy-Brändli/I. Oehri, Das Training künstlicher Intelligenz, [sic! 2023, 655 ff.](#); V. Salvadé, Comment licencier l'utilisation d'œuvres préexistantes pour entraîner l'intelligence artificielle?, [sic! 2024, 87 ss.](#); E. W. Brem/E. Brem, Künstliche Intelligenz und künstlerische Darbietung, [sic! 2024, 407 ff.](#), 417; C. Marti, Öffentliches Zurverfügungstellen eines Datensatzes zum Training von KI – Wie würden Schweizer Gerichte entscheiden?, [sic! 2025, 145 ff.](#); für Deutschland: T. Dornis/S. Stober, Urheberrecht und Training generativer KI-Modelle: Technologische und juristische Grundlagen, Baden-Baden 2024; M. Beurskens, Training generativer KI nur auf Lizenzgrundlage?, *RDi* 2025, 1 ff.; S. Ventroni, KI-Training als neue urheberrechtliche Nutzungsart: Erfassen Rechteklauseln in Lizenzverträgen das Recht zur Verwendung urheberrechtlich geschützter Inhalte für KI-Training?, *MMR* 2025, 163 ff.; für die USA: J. Quang, Does Training AI Violate Copyright Law?, *Berkeley Technology Law Journal*, Vol. 36, 1407 ff.; M. Sag/P. Yu, The Globalization of Copyright Exceptions for AI Training, *Emory Law Journal*, Vol. 74, (forthcoming 2025); abrufbar unter: papers.ssrn.com/sol3/papers.cfm?abstract_id=4976393# (20. Mai 2025); für das Vereinigte Königreich: A. Buick, Copyright and AI training data—transparency to the rescue?, *Journal of Intellectual Property Law & Practice*, Vol. 20, No. 3. März 2025, 182 ff.; E. Rosati, Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law, in: *European Journal of Risk Regulation*, 2024, 1 ff.

⁷ Hängig sind derzeit namentlich die folgenden Verfahren: *Getty Images (US), Inc. v. Stability AI, Inc.*, No. 1:23-CV-00135 (D. Del. 2023) [Verfahren in der Rechtsmittelinstanz]; *Sarah Andersen, et al. v. Stability AI Ltd., et al.*, No. 23-cv-00201-WHO, (N.D. Cal. Oct. 30, 2023); US District Court Southern District of New York, eingereicht am 27.12.23 (Case 1:23-cv-11195) – «New York Times vs. OpenAI & Microsoft» jetzt zusammengelegt mit Fall: US District Court Southern District of New York, eingereicht am 4. April 2025 (Case 1:23-cv-11195-SHS-OTW) – «New York Times and Daily Mail vs. OpenAI & Microsoft»; US District Court Northern District of California, eingereicht am 7. Juli 2023 (Case 3:23-cv-03416) – «Sarah Silverman et al. vs. OpenAI und Meta»; US District Court Southern District of New York, eingereicht am 24.6.2024 (Case 1:24-cv-04777) – «RIAA & Musiklabels vs. Udio AI»; US District Court District of Massachusetts, eingereicht am 24. Juni 2024 (Case 1:24-cv-11611) – «RIAA & Musiklabels vs. Suno AI»; US District Court Southern District of New York, eingereicht am 30.4.2024 (Case 1:24-cv-03285) – «Tribune Publishing vs. OpenAI & Microsoft»; US District Court Northern District of California, eingereicht am 28. Juni 2023 (Case 3:23-cv-03223).

⁸ D. Del. vom 11. Februar 2025, 2025 WL 458520, «Reuters v Ross».

⁹ [LG Hamburg](#) vom 27. September 2024, 310 O 227/23, ZUM 2025, 64, «LAION e.V.». (Berufung anhängig beim HansOLG, 5 U 111/20).

¹⁰ Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG (DSM-RL).

¹¹ Für Deutschland: §§ 44b und 60d UrhG; eingefügt durch das Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes vom 31. Mai 2021 (BGBl. I 1204); für Frankreich: Art. L122-5-3 Code de la propriété intellectuelle (CPI), eingeführt durch die Ordonnance n° 2021–1518 vom 24. November 2021.

¹² D. Rosenthal/L. Veraldi, Das Training von KI-Sprachmodellen mit fremden Inhalten und Daten aus rechtlicher Sicht, *Jusletter* vom 3. Februar 2025, N 36; Semmelmann (Fn. 3), 640; M. Städeli/L. Mary, Künstliche Intelligenz und Urheberrecht, [SJJ 2024, 244 ff.](#), 248 f.; P. Kübler, Wie generative KI-Systeme Rechte nutzen, *medialex* 05/2023, N 17; Cherpillod (Fn. 3), 447; Marmy-Brändli/Oehri (Fn. 6), 662 ff.

¹³ Für das deutsche Recht einzig: A. Wietzke, Retrieval Augmented Generation – eine urheberrechtliche Betrachtung, *RDi* 2024, 538 ff.

¹⁴ Motion Gössi 24.4596 «Besserer Schutz des geistigen Eigentums vor KI-Missbrauch» vom 20. Dezember 2024.

Wahrscheinlichkeitsrechnungen beruhen, ist es systemimmanent, dass ein gewisser Anteil der von diesen Modellen generierten Aussagen nicht zutrifft.

Das Halluzinieren kann zwar durch verschiedene Ansätze verringert, aber (bisher) nicht vollständig verhindert werden. Namentlich können grosse Sprachmodelle zusätzlich mit verlässlichen und kuratierten Daten trainiert werden, um die Genauigkeit ihrer Aussagen zu verbessern (sog. *fine tuning*). Auch spezialisierte Submodelle oder *Fact-Checking*-Algorithmen können hilfreich sein. Bei heiklen Anwendungen (bspw. im Gesundheitsbereich) empfiehlt es sich zudem, dass der Output von einem Menschen überprüft wird. Ein besonders erfolgversprechender und bereits weit verbreiteter Ansatz besteht sodann darin, dass Modelle auf Datenbanken oder das Internet zugreifen, um die Zuverlässigkeit und Aktualität des Outputs zu verbessern. Dieser Ansatz wird als *Retrieval-Augmented Generation* ([RAG](#)) bezeichnet. Das Modell kombiniert dabei sein allgemeines «Wissen» und seine Fähigkeit zur Formulierung von Antworten mit spezifischen Informationen, die in einer Datenbank eines Unternehmens gespeichert oder im Internet abrufbar sind. Beispiele sind KI-Assistenten für Anwaltskanzleien, die durch Zugriff auf eine interne Datenbank mit von der Kanzlei verfassten Verträgen in der Lage sind, qualitativ hochwertige Entwürfe für weitere Verträge zu erstellen, oder Chatbots von Versicherungen, die durch Zugriff auf relevante Dokumente (bspw. interne Richtlinien oder AGB) Fragen von Kunden zutreffend beantworten können. Zu denken ist aber auch an Dienste, die über das Internet auf die Inhalte von Medien zugreifen und ihre Nutzer damit tagesaktuell und zuverlässig über das aktuelle Geschehen informieren können, bspw. der Dienst von Perplexity.ai.

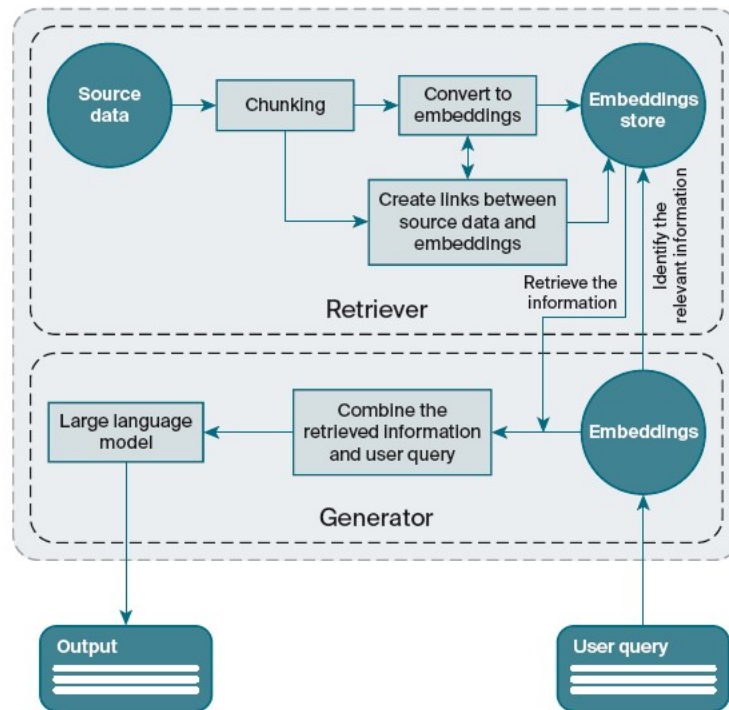
Die Verwendung von [RAG](#) wirft eine Reihe von urheberrechtlichen Fragen auf, die hier untersucht werden. Im Sinne eines *caveat* ist allerdings darauf hinzuweisen, dass sich dieser Aufsatz nicht als Beitrag zum rechtswissenschaftlichen Diskurs versteht. Vielmehr wird – im Sinne einer klassischen Subsumtion – ein konkretes Phänomen urheberrechtlich eingeordnet und es werden erste Leitlinien für die urheberrechtliche Beurteilung skizziert.

II. Retrieval-Augmented Generation ([RAG](#))

Grosse Sprachmodelle wie GPT 4.o, Gemini, Mistral Large oder DeepSeek-V3 werden mit riesigen Datenmengen trainiert. Da die Kosten für das Training sehr hoch sind, können diese Modelle nicht laufend mit aktuellen Daten nachtrainiert und aktualisiert werden. Die in Sprachmodellen enthaltenen Daten sind deshalb rasch veraltet. Hier setzt [RAG](#) an, indem es zwei Komponenten miteinander verbindet: Beim sog. *Retrieval* wird eine interne Datenbank oder das Internet verwendet, um für das Beantworten einer Anfrage relevante Informationen zu identifizieren und abzurufen. Bei der sog. *Generation* verarbeitet das Sprachmodell die abgerufenen Informationen und generiert auf deren Grundlage eine (in der Regel) präzise und zutreffende Antwort.

Die Funktionsweise von [RAG](#) lässt sich im Wesentlichen wie folgt grafisch darstellen:¹⁵

¹⁵ A. Lawrence, Understanding [RAG](#): Retrieval-Augmented Generation Explained, 2. Februar 2024, <www.6clicks.com/resources/blog/understanding-rag-retrieval-augmented-generation-explained> (20. Mai 2025).



(1) Der Prozess beginnt mit der Eingabe eines Prompts durch einen Nutzer (*User query*). Diese Eingabe wird als Suchanfrage interpretiert und vom System verarbeitet. Dabei wird der eingegebene Text analysiert, um relevante Schlüsselbegriffe und semantische Beziehungen zu identifizieren. Das erfolgt mithilfe sog. *Embeddings*. Dabei handelt es sich um numerische Vektordarstellungen von Texten oder anderen Daten, die es einem KI-Modell erlauben, Ähnlichkeiten mit den Inhalten der abzufragenden Datenbank zu erkennen. Dies ermöglicht eine semantische Suche, bei der nicht nur nach exakten Schlüsselwörtern, sondern auch nach inhaltlich verwandten Informationen gesucht wird.

(2) Im *Retriever* werden die relevanten Informationen bereitgestellt, die für die Beantwortung des Prompts gebraucht werden. Der *Retriever*-Prozess beginnt mit der Vorbereitung der *Source data* (Quell-Daten) und endet mit der Speicherung der generierten *Embeddings* in einem Vektorspeicher (*Embeddings store*). Die Daten können aus verschiedenen Quellen stammen und in verschiedenen Formaten vorliegen, Texte bspw. als PDF oder als HTML-Code. Die Daten werden verarbeitet, indem sie in kleinere Teile, bspw.

sic! 2025 S. 381, 384

Absätze eines Textes, aufgeteilt (*chunking*) und anschliessend in Vektordarstellungen umgewandelt (*Embedding*) und mit der *Source data* verlinkt werden. Die *Embeddings* werden sodann im *Embeddings store* gespeichert und mit den *Embeddings* der *User query* verglichen, um die relevanten Informationen in der *Source data*, bspw. ein (oder mehrere) Dokument(e), zu identifizieren. Weiterverarbeitet werden dabei meist nur die besten Treffer (häufig die besten fünf oder zehn); so können die relevantesten Dokumente für die Texterzeugung priorisiert werden. Diese werden dann für die Generierung der Antwort an den *Generator* weitergegeben.

(3) Der *Generator* erstellt aufgrund der erhaltenen Dokumente eine Antwort auf die *User query*. Dabei werden die *User query* und die Dokumente vom System verbunden und dem Sprachmodell als Prompt eingegeben. Dieses analysiert die bereitgestellten Informationen und generiert auf dieser Basis eine Antwort (*Output*). Die Dokumente können dabei direkt und vollständig in den Prompt integriert werden; bei umfangreichen Dokumenten werden stattdessen relevante Abschnitte zusammengefasst und dem Sprachmodell in komprimierter Form zur Verfügung gestellt. Beim Erzeugen der Antwort (*Output*) nutzt das Sprachmodell linguistische Techniken, um klare und verständliche Formulierungen zu generieren. Zusätzlich kann ein Validierungsmechanismus integriert werden, um die Richtigkeit und Konsistenz der Antwort zu überprüfen. Auch hier gibt es verschiedene Ansätze: Das Sprachmodell kann anhand interner Wahrscheinlichkeiten bewerten, ob die Antwort plausibel ist; alternativ lassen sich externe *Fact-Checking*-Mechanismen verwenden, bspw. spezielle KI-Modelle oder *Application Programming Interfaces (APIs)*, um die generierte Antwort mit vertrauenswürdigen Quellen abzugleichen. Zudem können unklare oder fehlerhafte Antworten aufgrund von Rückmeldungen der Nutzer erkannt und künftige Generationen des Sprachmodells verbessert werden.

III. RAG mit Zugriff auf interne Datenbank

Bei der «klassischen» Nutzung von RAG greift das Sprachmodell auf die Inhalte einer internen Datenbank zu, also auf eine Datenbank, die von einem Unternehmen oder einer anderen Organisation selbst betrieben wird. Aus Sicht des Urheberrechts lassen sich die Vorgänge bei dieser Form von RAG in drei Schritte unterteilen: (1) das Speichern von Werken in der Datenbank, (2) der Zugriff auf diese Werke, die Identifikation der relevanten Inhalte und deren Integration in einen Prompt sowie (3) die allfällige Wiedergabe von Werken oder Teilen davon in der Ausgabe (Output).

1. Speichern von Werken

Das Speichern von Werken in einer internen Datenbank ist keine RAG-spezifische Frage. Die damit verbundenen Vervielfältigungen ([Art. 10 Abs. 2 lit. a URG](#)) sind zulässig, wenn die Rechteinhaber zustimmen oder eine Schranke greift. Das gilt für das Speichern der Werke als *Source data* ebenso wie für das Speichern als numerische Vektordarstellungen im *Embeddings store*. Denn die Art der Vervielfältigung spielt urheberrechtlich keine Rolle.¹⁶ Für das Umwandeln der Werke in Vektordarstellungen und die Verlinkung mit der *Source data* ist die Zustimmung der Rechteinhaber nicht erforderlich, weil die bloße Änderung des technischen Formats den individuellen Charakter des Werks nicht berührt und damit weder eine Umgestaltung noch eine Bearbeitung vorliegt ([Art. 11 Abs. 1 URG](#)).¹⁷

In Datenbanken von Unternehmen werden oft Werke gespeichert, die von *Mitarbeitenden* erstellt werden. Soweit sich die Arbeitgeberin die Urheberrechte an diesen Werken übertragen liess, kann sie über deren Speichern in einer Datenbank entscheiden. Werke Dritter dürfen in Datenbanken von Unternehmen und anderen Organisationen auch ohne Zustimmung der Rechteinhaber gespeichert werden, wenn ein *interner Gebrauch* ([Art. 19 Abs. 1 lit. c URG](#)) vorliegt. Diese Schranke stellt allerdings nur Vervielfältigungen frei, die der *Information oder Dokumentation* innerhalb eines Betriebes, einer öffentlichen Verwaltung, eines Instituts, einer Kommission oder einer ähnlichen Einrichtung dienen. Die Freistellung erfasst damit nur die Vervielfältigung von Werkexemplaren, die einen inhaltlichen Bezug zur Tätigkeit des Betriebs (etc.) aufweisen. Das wird hier regelmässig der Fall sein, zumal der Begriff der Information oder Dokumentation weit zu verstehen ist.¹⁸ Nicht freigestellt ist allerdings die *vollständige oder weitgehend vollständige Vervielfältigung* im Handel erhältlicher Werkexemplare ([Art. 19 Abs. 3 lit. a URG](#)). Diese Gegen Ausnahme schränkt das Erstellen umfassender interner Datenbanken massgeblich ein. Sollen Werke Dritter vollständig in einer internen Datenbank gespeichert werden, muss deshalb die Zustimmung der Rechteinhaber eingeholt werden. Soweit Vervielfältigungen als interner Gebrauch freigestellt sind, ist dafür eine *Vergütung* geschuldet ([Art. 20 Abs. 2 URG](#)). Diese wird heute im Rahmen des GT 8 bezahlt.

2. Zugriff auf Werke, Identifikation relevanter Inhalte und Integration in einen Prompt

Beim Zugriff auf die in einer internen Datenbank gespeicherten Werke, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt werden Vervielfältigungen erstellt.

sic! 2025 S. 381, 385

¹⁶ M. Reh binder/L. Haas/K-P. Uhlig, [URG](#) Kommentar, 4. Aufl., Zürich 2022, [URG 10](#) N 10; R. M. Hilty, Urheberrecht, 2. Aufl., Bern 2020, N 303 f.

¹⁷ Ebenso: D. Barrelet/W. Egloff, in: D. Barrelet/W. Egloff (Hg.), Das neue Urheberrecht, 4. Aufl., Bern 2020, [URG 11](#) N 8; H. Pfortmüller, in: B. Müller/R. Oertli (Hg.), Stämpfli Handkommentar zum Urheberrechtsgesetz ([URG](#)), 2. Aufl., Bern 2012, [URG 11](#) N 4; Hilty (Fn. 16), N 198; Reh binder/Haas/Uhlig (Fn. 16), [URG 11](#) N 6; Y. Benhamou/A. Andrijevic, Intelligence artificielle générative d'images et droit d'auteur, in: Aspects juridiques de l'intelligence artificielle, Bern 2024, 41 ss, 53.

¹⁸ Reh binder/Haas/Uhlig (Fn. 16), [URG 19](#) N 30; Barrelet/Egloff (Fn. 17), [URG 19](#) N 20; C. Gasser, in: B. Müller/R. Oertli (Hg.), Stämpfli Handkommentar zum Urheberrechtsgesetz ([URG](#)), 2. Aufl., Bern 2012, [URG 19](#) N 19.

Soweit Werke Dritter ohne deren Zustimmung genutzt werden, stellt sich die Frage, ob das Erstellen der Vervielfältigungen durch eine Schranke freigestellt wird. In Frage kommen die «Schranke»¹⁹ für vorübergehende Vervielfältigungen ([Art. 24a URG](#)) und die Schranke zugunsten des internen Gebrauchs ([Art. 19 Abs. 1 lit. c URG](#)), nicht aber die Wissenschaftsschranke ([Art. 24d URG](#)). Denn anders als bei der Nutzung von Werken für das Training von KI-Modellen, die sich allenfalls als Nutzung zum Zweck der wissenschaftlichen Forschung verstehen lässt,²⁰ werden Werke bei der Verwendung von [RAG](#) nicht für die wissenschaftliche Forschung – nämlich für die Entwicklung eines Modells – genutzt, sondern für die Beantwortung einer Anfrage bei der Verwendung eines trainierten Modells. Die Wissenschaftsschranke greift deshalb nur, wenn [RAG](#) als Mittel der wissenschaftlichen Forschung verwendet wird, bspw. für die Identifikation der weiblichen Figuren in den Werken von Thomas Mann.

a) Vorübergehende Vervielfältigungen

Die vorübergehende Vervielfältigung von Werken ist zulässig, wenn vier Voraussetzungen erfüllt sind: Die Vervielfältigung muss flüchtig oder begleitend sein ([Art. 24a lit. a URG](#)); sie muss einen integralen und wesentlichen Teil eines technischen Verfahrens darstellen ([Art. 24a lit. b URG](#)); sie muss ausschliesslich der Übertragung in einem Netz zwischen Dritten durch einen Vermittler oder einer rechtmässigen Nutzung dienen ([Art. 24a lit. c URG](#)); und sie darf keine eigenständige wirtschaftliche Bedeutung haben ([Art. 24a lit. d URG](#)).

aa) Flüchtig oder begleitend

Die erste Voraussetzung umfasst zwei alternative Merkmale, die Vervielfältigung muss also entweder flüchtig oder begleitend sein ([Art. 24a lit. a URG](#)). Flüchtig sind Vervielfältigungen, die besonders kurzlebig sind, also mit der Beendigung des technischen Vorgangs gleich wieder gelöscht werden.²¹ Begleitend sind Vervielfältigungen, die beiläufig im Rahmen eines technischen Verfahrens erstellt werden und damit der Erleichterung eines anderen Nutzungsvorgangs dienen.²² Die Vervielfältigungen, die beim Zugriff auf die Werke in einer Datenbank, bei der Identifikation der relevanten Inhalte und bei der Integration in einen Prompt erstellt werden, werden nach diesem Vorgang gleich wieder gelöscht und sind damit flüchtig.

bb) Integraler und wesentlicher Teil eines technischen Verfahrens

Nach der zweiten Voraussetzung muss die Vervielfältigung einen integralen und wesentlichen Teil eines technischen Verfahrens darstellen ([Art. 24a lit. b URG](#)). Das ist der Fall, wenn das technische Verfahren massgeblich auf der Vervielfältigung von Werken beruht.²³ Die Vervielfältigung muss aber nicht technisch strikt notwendig sein; vielmehr reicht es, wenn sie massgeblich zum effizienten Funktionieren eines technischen Verfahrens beiträgt.²⁴ Auch diese Voraussetzung ist erfüllt, weil [RAG](#) als technisches Verfahren nur funktioniert, wenn das System auf die Werke in einer Datenbank zugreifen, die relevanten Inhalte identifizieren und in einen Prompt integrieren kann.

¹⁹ Zur umstrittenen Frage, ob [Art. 24a URG](#) als Schranke oder als Konkretisierung des Begriffs der Vervielfältigung zu verstehen ist, siehe: Hilty (Fn. 16), N 473, der die Bestimmung als Konkretisierung des Vervielfältigungsbegriffs versteht; ebenso F. Thouvenin, Immaterialgüterrecht, Zürich 2025, N 621; für dieses Verständnis spricht auch das Vorbild im EU-Recht, zumal Art. 5 Abs. 1 DSM-RL nicht von einer Beschränkung spricht, sondern die Formulierung verwendet: «werden von dem in Artikel 2 vorgesehenen Vervielfältigungsrecht ausgenommen»; als Schranke qualifizierend dagegen: Botschaft URG 2006, BBl 2006, 3389 ff., 3403 und 3430 f.; Oertli (Fn. 18), [URG 24a](#) N 3.

²⁰ Siehe dazu: Rosenthal/Veraldi (Fn. 12), N 36, die KI-Modelle per Definition als Erkenntnisse verstehen, weil diese in Form ihrer Parameter das Ergebnis der Analyse der Trainingsinhalte verkörpern und anhand von generalisierten Regeln Auskunft darüber geben, wie Sprache verwendet werde, um gestützt darauf neue Texte generieren zu können; so wohl auch Marti (Fn. 6), 151 f., mit Bezug auf den Sachverhalt des Urteils des [LG Hamburg](#) vom 27. September 2024 (310 O 227/23); a.M. wohl Städeli/Mary (Fn. 12), welche die Anwendbarkeit der Wissenschaftsschranke nur bejahen, wenn das Training generativer KI-Modelle ausschliesslich für die wissenschaftliche Forschung erfolgt; zurückhaltend Thouvenin/Picht (Fn. 2), 515.

²¹ Rehbinders/Hasas/Uhlig (Fn. 16), [URG 24a](#) N 4; Egloff (Fn. 18), [URG 24a](#) N 4; Hilty (Fn. 16), N 471; Oertli (Fn. 18), [URG 24a](#) N 5; P. Gilliéron, in: J. de Werra/P. Gilliéron (Hg.), *Propriété intellectuelle*, Bâle 2013, LDA 24a N 12.

²² Thouvenin (Fn. 19), N 620; Rehbinders/Hasas/Uhlig (Fn. 16), [URG 24a](#) N 4; Egloff (Fn. 18), [URG 24a](#) N 4.

²³ Rehbinders/Hasas/Uhlig (Fn. 16), [URG 24a](#) N 5; Egloff (Fn. 18), [URG 24a](#) N 5; Gilliéron (Fn. 21), LDA 24a N 16.

²⁴ Rehbinders/Hasas/Uhlig (Fn. 16), [URG 24a](#) N 5; Egloff (Fn. 18), [URG 24a](#) N 5.

cc) Übertragung in einem Netz oder rechtmässige Nutzung

Bei der dritten Voraussetzung sind zwei Konstellationen zu unterscheiden: Die Vervielfältigung muss entweder ausschliesslich der Übertragung in einem Netz zwischen Dritten durch einen Vermittler oder einer rechtmässigen Nutzung dienen ([Art. 24a lit. c URG](#)). Rechtmässig sind Nutzungen, welche der Rechteinhaber erlaubt hat, und solche, die gesetzlich zulässig sind, die also keinem Ausschliesslichkeitsrecht unterliegen oder von einer *Schranke* freigestellt werden.²⁵ Eine rechtmässige Nutzung im Sinn dieser Bestimmung ist auch der *Werkgenuss*, also die Wahrnehmung von Werken mit den menschlichen Sinnen, der urheberrechtlich frei ist.²⁶ Freigestellt sind damit auch vorübergehende Vervielfältigungen, die unmittelbar dem Werkgenuss dienen, bspw. das Anzeigen von Inhalten auf einem Bildschirm, das Zwischenspeichern beim Abspielen einer DVD oder das

sic! 2025 S. 381, 386

Speichern von Inhalten im Arbeitsspeicher beim Abrufen einer Webseite (*Browsing*).²⁷

Die erste Konstellation zielt auf Vervielfältigungen beim *Übermitteln von Werken über das Internet*. Eine Übertragung in einem Netz zwischen Dritten durch einen Vermittler liegt hier aber nicht vor. Zu prüfen ist damit, ob die Vervielfältigungen, die beim Zugriff auf die Werke in einer Datenbank, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt erstellt werden, im Sinn der zweiten Konstellation einer *rechtmässigen Nutzung* dienen. Diese Vervielfältigungen dienen zwar letztlich dem Werkgenuss, also der Wahrnehmung der in den Werken enthaltenen Informationen mit menschlichen Sinnen, dies aber nur mittelbar, indem den Werken die relevanten Informationen entnommen und beim Erzeugen des Outputs verwendet werden. Mit den menschlichen Sinnen wahrgenommen wird nur der mit Hilfe von [RAG](#) erstellte Output von Sprachmodellen, nicht die Werke, die bei der Generierung des Outputs verwendet werden. Die hier in Frage stehenden Vervielfältigungen dienen damit nicht unmittelbar dem Werkgenuss.

Einer rechtmässigen Nutzung dienen die Vervielfältigungen aber, wenn [RAG](#) für Zwecke der internen Information oder Dokumentation verwendet wird, weil die dafür erstellten Vervielfältigungen in diesem Fall von der Schranke zugunsten des *internen Gebrauchs* ([Art. 19 Abs. 1 lit. c URG](#)) freigestellt werden. Wie stets greift auch hier die Gegen Ausnahme, nach der im Handel erhältliche Werkexemplare nicht vollständig oder weitgehend vollständig vervielfältigt werden dürfen ([Art. 19 Abs. 3 lit. a URG](#)). Die dritte Voraussetzung für die Zulässigkeit vorübergehender Vervielfältigungen ist damit nur erfüllt, wenn die Werke mit Zustimmung der Rechteinhaber oder im Rahmen des internen Gebrauchs verwendet werden.²⁸

dd) Keine eigenständige wirtschaftliche Bedeutung

Nach der vierten Voraussetzung darf die Vervielfältigung keine eigenständige wirtschaftliche Bedeutung haben ([Art. 24a lit. d URG](#)). Dieses Erfordernis spiegelt nach einem Teil der Lehre²⁹ die Vorgaben der zweiten und dritten Stufe des Dreistufentests, nach der Schranken die normale Verwertung eines Werks nicht beeinträchtigen und die berechtigten Interessen der Rechteinhaber nicht unangemessen verletzen dürfen.³⁰ Da die Bestimmung von [Art. 24a URG](#) richtigerweise nicht als Schranke, sondern als Konkretisierung des Begriffs der Vervielfältigung zu qualifizieren ist,³¹ kann dieser Auffassung zwar nicht ohne Weiteres gefolgt werden. Die inhaltlichen Bezüge zwischen der zweiten und dritten Stufe des Dreistufentests und dem Begriff der eigenständigen wirtschaftlichen Bedeutung sind aber so eng, dass sich die Auslegung dieses Tatbestandsmerkmals dennoch am Verständnis dieser beiden Stufen des Dreistufentests orientieren kann.

²⁵ Thouvenin (Fn. 19), N 620; Reh binder/Haas/Uhlig (Fn. 16), [URG 24a](#) N 8, jedoch ohne Erwähnung der Schranken; Egloff (Fn. 18), [URG 24a](#) N 6.

²⁶ Thouvenin (Fn. 19), N 613; Reh binder/Haas/Uhlig (Fn. 16), [URG 10](#) N 3; Barrelet/Egloff (Fn. 17), [URG 10](#) N 8; Hilty (Fn. 16), N 292 ff.; Cherpillod (Fn. 21), LDA 10 N 3.

²⁷ [BGE 133 III 473 ff. E. 4.4](#) «Pressespiegel»; Botschaft URG 2006, BBl 2006, 3389 ff., 3430 f. Reh binder/Haas/Uhlig (Fn. 16), [URG 24a](#) N 8.

²⁸ Siehe dazu nachfolgend, III.2.b).

²⁹ Egloff (Fn. 18), [URG 24a](#) N 7; Oertli (Fn. 18), [URG 24a](#) N 12; Gilliéron (Fn. 21), LDA 24a N 21.

³⁰ Art. 13 TRIPS; Art. 9 Abs. 2 RBÜ; [Art. 10 WCT](#); [Art. 16 WPPT](#).

³¹ Siehe dazu vorne, Fn. 19.

Eine eigenständige wirtschaftliche Bedeutung hat eine Vervielfältigung, wenn sie erlaubt, Werke anders oder intensiver zu nutzen, als dies ohne Vervielfältigung möglich wäre, wenn sie also eine *neue, eigenständige Nutzungsmöglichkeit* eröffnet.³² Der wirtschaftliche Nutzen einer Vervielfältigung darf mithin nicht über den Vorteil hinausgehen, der sich aus der Übertragung in einem Netz oder aus der rechtmässigen Nutzung des Werks ergibt.³³ Ob diese Voraussetzung hier erfüllt ist, erscheint zweifelhaft. Zwar lässt sich durchaus argumentieren, dass die im Rahmen von [RAG](#) erstellten Vervielfältigungen keine Nutzung der Werke ermöglichen, die nicht auch ohne diese Vervielfältigungen möglich wäre, zumal der Unterschied allein in der Reduktion des Aufwandes liegt, der bei einer menschlichen Identifikation relevanter Dokumente und Inhalte ungleich grösser ist als beim Einsatz eines Retrievers.³⁴ Andererseits erlaubt [RAG](#) gerade wegen der geringeren Kosten und der höheren Geschwindigkeit der Identifikation und Entnahme der relevanten Inhalte, Werke anders und intensiver zu nutzen als dies bisher möglich war.

Die normale Verwertung von Werken wird durch [RAG](#) zwar nicht beeinträchtigt, wenn man die Betrachtung auf die bisherigen Verwertungsformen beschränkt. Bei der Beurteilung der zweiten Stufe des Dreistufentests ist aber anerkannt, dass auch *künftige Verwertungsformen* einzubeziehen sind, wenn sie nicht rein hypothetisch erscheinen, sondern aufgrund der technischen und wirtschaftlichen Entwicklung absehbar sind.³⁵ Diese Perspektive erscheint auch beim Tatbestandsmerkmal der eigenständigen wirtschaftlichen Bedeutung angemessen. Angesichts der schon heute grossen und wohl weiter zunehmenden Bedeutung von [RAG](#) ist durchaus denkbar, dass für das Recht zur Nutzung von Werken im Rahmen von [RAG](#) künftig Lizenzen erteilt werden, wenn die Nutzung nicht gesetzlich freigestellt ist. Eine weite Auslegung der Schranke von [Art. 24a URG](#) würde das Entstehen eines solchen Marktes und die damit verbundene Möglichkeit zur Verwertung von Urheberrechten verhindern. Auch wenn die Frage nicht abschliessend beurteilt werden kann, wird man wohl davon ausgehen müssen, dass die vorübergehenden Vervielfältigungen, die im Rahmen von [RAG](#)

sic! 2025 S. 381, 387

erstellt werden, eine eigenständige wirtschaftliche Bedeutung haben oder zumindest künftig haben könnten.

ee) Fazit

Trotz gewisser Unsicherheiten ist anzunehmen, dass die Vervielfältigungen, die beim Zugriff auf die Werke in einer Datenbank, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt erstellt werden, die Voraussetzungen für die Qualifikation als vorübergehende Vervielfältigung im Sinn von [Art. 24a URG](#) in der Regel nicht erfüllen.

b) Interner Gebrauch

Ob die Verwendung von Werken im Rahmen von [RAG](#) als interner Gebrauch freigestellt ist, wurde bereits für die Speicherung von Werken in einer internen Datenbank geprüft und bejaht.³⁶ Für die hier in Frage stehenden Vervielfältigungen gilt nichts anderes. Wie bei der Speicherung in einer internen Datenbank bestehen allerdings auch hier relevante Einschränkungen, weil im Handel erhältliche Werkexemplare nicht vollständig oder weitgehend vollständig vervielfältigt werden dürfen ([Art. 19 Abs. 3 lit. a URG](#)).³⁷ Ist das nicht der Fall, sind auch die Vervielfältigungen, die beim Zugriff auf die Werke in einer internen Datenbank, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt erstellt werden, als zulässiger interner Gebrauch zu qualifizieren. Für diese Nutzungen ist eine *Vergütung* geschuldet ([Art. 20 Abs. 2 URG](#)), die nur von einer zugelassenen Verwertungsgesellschaft geltend gemacht werden kann ([Art. 20 Abs. 4 URG](#)). Diese Nutzungen werden bei der Verhandlung künftiger Tarife zu berücksichtigen sein.

³² I. Cherpillod, SIWR II/1, 4. Aufl., Basel 2024, N 886; Reh binder/Haas/Uhlig (Fn. 16), [URG 24a](#) N 9; Egloff (Fn. 18), [URG 24a](#) N 7; Oertli (Fn. 18), [URG 24a](#) N 12, der von einer «quantitativ oder qualitativ andere[n] Nutzung» spricht; nach Gilliéron (Fn. 21), LDA 24a N 22 fehlt eine eigenständige wirtschaftliche Bedeutung, wenn die Vervielfältigung keinerlei unabhängige wirtschaftliche Bedeutung hat.

³³ Zu Art. 5 Abs. 1 der Richtlinie 2001/29/EG zur Harmonisierung bestimmter Aspekte des Urheberrechts und der verwandten Schutzrechte in der Informationsgesellschaft: EuGH vom 4. Oktober 2011, C-403/08 und C-429/08, N 175.

³⁴ So für das deutsche (und europäische) Recht: F. Hofmann, Retten Schranken Geschäftsmodelle generativer KI-Systeme?, ZUM 2024, 166, 169; a.M.: A. Wietzke (Fn. 1314), N 34.

³⁵ S. Ricketson/J. Ginsburg, International Copyrights and Neighbouring Rights, 3. Aufl., Oxford, 2022, Rz. 13.16; WTO Panel Report vom 15. Juni 2000, WT/DS160/R – United States Section 110(5) of the US Copyright Act, Rz. 6.166 ff.

³⁶ Siehe dazu vorne, III.1.

³⁷ Siehe dazu vorne, III.1.

3. Werke im Output

Der Output von Sprachmodellen kann Teile von Werken enthalten oder bestehenden Werken so ähnlich sein, dass er in deren Schutzbereich fällt. In der Regel wird das allerdings nicht der Fall sein, weil Sprachmodelle – auch bei der Nutzung von [RAG](#) – nicht Texte duplizieren, sondern neu erstellen.

Eine Urheberrechtsverletzung liegt allerdings selbst bei einem Eingreifen in den Schutzbereich nur vor, wenn ein *bestehendes Werk* bei der Erzeugung eines neuen Werks *tatsächlich genutzt* worden ist.³⁸ Haben mehrere Personen unabhängig voneinander identische oder weitgehend identische Werke geschaffen, liegt eine *Doppel- oder Parallelschöpfung* vor; in diesem Fall verletzt das jüngere Werk die Urheberrechte am älteren Werk nicht, weil Letzteres bei der Schöpfung des Ersteren nicht verwendet wurde.³⁹ Dieser Fall ist zwar äusserst selten – jedenfalls solange Werke (nur) von Menschen geschaffen werden. Es ist aber denkbar, dass die Verwendung von generativer KI dazu führen wird, dass der Output von KI-Systemen bestehenden Werken häufiger so ähnlich sein wird, dass er in deren Schutzbereich fällt. Eine Urheberrechtsverletzung liegt aber auch in diesen Fällen nur bei der *Verwendung eines geschützten Werks* vor. Das ist bei Sprachmodellen nur der Fall, wenn ein *Werk beim Training des Modells oder bei dessen Einsatz verwendet* wurde, im Fall von [RAG](#) mit Zugriff auf eine interne Datenbank also, wenn das Werk in der internen Datenbank gespeichert und der Output unter Zugriff auf dieses Werk generiert worden ist. Trifft das nicht zu, verletzt der Output eines Sprachmodells selbst dann keine Urheberrechte, wenn er mit einem bestehenden Werk identisch oder diesem so ähnlich ist, dass er in dessen Schutzbereich fällt.

Enthält der Output für sich geschützte Teile von Werken oder ist er einem bestehenden Werk so ähnlich, dass er in dessen Schutzbereich fällt, und wurde das Werk (oder Teile davon) beim Training oder beim Einsatz eines KI-Systems tatsächlich verwendet, stellt sich die Frage, ob diese Nutzung durch eine Schranke freigestellt ist. Wird der Output nur für interne Zwecke genutzt, wird die Vervielfältigung von Werken im Output von der Schranke zugunsten des *internen Gebrauchs* ([Art. 19 Abs. 1 lit. c URG](#)) erfasst. Fraglich ist allerdings, ob diese Schranke nur unveränderte Vervielfältigungen von Werken oder auch Vervielfältigungen in veränderter Form freistellt. Diese Frage wurde in der Lehre kaum untersucht und erscheint offen.⁴⁰ Sie kann hier zwar nicht abschliessend geklärt werden, bei einer kursorischen Betrachtung spricht aber einiges dafür, den Begriff der Vervielfältigung in [Art. 19 Abs. 1 lit. c URG](#) so zu verstehen, dass er *Vervielfältigungen von Werken in unveränderter und veränderter Form* erfasst. Dafür spricht schon der Wortlaut, der nicht zwischen verschiedenen Formen der Vervielfältigung unterscheidet. Hinzu kommt, dass [Art. 19 Abs. 1 lit. c URG](#) weit zu verstehen ist. Auch wenn das Gesetz nur die Vervielfältigung erwähnt, ist allgemein anerkannt, dass die Schranke auch weitere Nutzungen freistellt, namentlich das Verbreiten und Zugänglichmachen von Vervielfältigungen innerhalb einer Organisation.⁴¹ Für ein Verständnis, das veränderte und unveränderte Vervielfältigungen umfasst, spricht auch die Verwendung des Begriffs an anderer Stelle

sic! 2025 S. 381, 388

in [Art. 19 URG](#), namentlich in [Art. 19 Abs. 3 URG](#). Hier ist der Begriff der Vervielfältigung zweifellos in diesem umfassenden Sinn zu verstehen. Das gegenteilige Verständnis würde dazu führen, dass Werke der bildenden Kunst, grafische Aufzeichnungen von Werken der Musik und im Handel erhältliche Werkexemplare vollständig oder weitgehend vollständig vervielfältigt werden dürften, wenn sie verändert worden sind; ausgeschlossen wären diese Handlungen nur bei unveränderten Vervielfältigungen. Ein solches Ergebnis liesse sich mit Sinn und Zweck der Gegenausnahmen nicht vereinbaren. Für ein umfassendes Verständnis des Begriffs der Vervielfältigung sprechen auch pragmatische Gründe: Würde man veränderte Vervielfältigungen im Rahmen des internen Gebrauchs nicht zulassen, könnten die Rechteinhaber den durch [Art. 19 Abs. 1 lit. c URG](#) privilegierten Organisationen das automatisierte Erstellen von Zusammenfassungen

³⁸ Reh binder/Haas/Uhlig (Fn. 16), [URG 2](#) N 14; Thouvenin (Fn. 19), N 672 f.; zum deutschen Recht: BGH vom 5. Juni 1970, I ZR 44/68, «Magdalenenarie»; OLG Zweibrücken vom 19. November 2015, 4 U 186/14, N 27, «Piano-Lehrbuch»; U. Loewenheim, in: Schricker/Loewenheim, 6. Aufl. 2020, UrhG § 23 N 34, 36; B. Raue, in: Dreier/Schulze, 8. Aufl. 2025, UrhG § 23 N 35.

³⁹ Reh binder/Haas/Uhlig (Fn. 16), [URG 2](#) N 13; Einzelheiten sind allerdings strittig; siehe zum Ganzen: E. Haas, Die Verwendung von Bearbeitungen urheberrechtlich geschützter Werke, Baden-Baden 2018, 105 ff.

⁴⁰ Die Lehre äussert sich zu dieser Frage nicht explizit, geht aber in der Regel davon aus, dass im Rahmen der Schranken zugunsten des Privatgebrauchs und des Schulgebrauchs eine Änderung von Werken zulässig ist, woraus sich *e contrario* der Schluss ziehen lässt, dass dies für den internen Gebrauch nicht gelten soll; siehe dazu: Barrelet/Egloff (Fn. 17), [URG 19](#) N 14 und N 19; Reh binder/Haas/Uhlig (Fn. 16), [URG 24c](#) N 6; C. Grasser, Der Eigengebrauch im Urheberrecht, Bern 1997, 66 f.; Botschaft URG 1989, BBl 1989 III 477 ff., 539.

⁴¹ Botschaft URG 1989, BBl 1989 III 477 ff., 540; Botschaft URG 2006, BBl 3389 ff., 3421; [BGE 133 III 473 ff. E. 3.1 f.](#) «Pressespiegel»; Barrelet/Egloff (Fn. 17), [URG 19](#) N 20; Gasser (Fn. 18), [URG 19](#) N 21; Cherpillod (Fn. 32), N 278.

von Texten (und anderen Werken) auch dann verbieten, wenn diese die dafür verwendeten Werkexemplare im Handel erworben haben. Zulässig wären Zusammenfassungen nur, wenn sie nicht in den Schutzbereich der zusammengefassten Werke fallen würden. Die komplexe Frage nach dem Eingriff in den Schutzbereich müsste jeweils im Einzelfall geprüft werden, wobei eine Prüfung erst möglich wäre, wenn eine Zusammenfassung erstellt und die Urheberrechte möglicherweise bereits verletzt worden sind. Ein solches Ergebnis erscheint wenig sinnvoll. Insgesamt spricht damit viel dafür, den Begriff der Vervielfältigung in [Art. 19 Abs. 1 lit. c URG](#) so zu verstehen, dass er unveränderte und veränderte Vervielfältigungen gleichermaßen umfasst.

Wie stets greift auch bei der Vervielfältigung von Werken im Output die Gegenausnahme, nach der im Handel erhältliche Werkexemplare nicht vollständig oder weitgehend vollständig vervielfältigt werden dürfen ([Art. 19 Abs. 3 lit. a URG](#)). Anders als beim Erstellen einer internen Datenbank und beim Zugriff auf die Werke, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt, wird der Output beim Einsatz von [RAG](#) allerdings kaum je vollständig oder weitgehend vollständig einem im Handel erhältlichen Werkexemplar entsprechen. Wäre dies der Fall, würde der Output also den Input weitgehend unverändert wiedergeben, könnte auf den Einsatz von [RAG](#) verzichtet und stattdessen das ursprüngliche Dokument verwendet werden. Soweit der Output nur intern genutzt wird, ergeben sich damit in aller Regel keine relevanten Einschränkungen.

Anderes gilt, wenn der *Output* nicht für interne Zwecke genutzt, sondern *Dritten zugänglich* gemacht wird, etwa bei einem Chatbot, der Fragen von Kunden beantwortet. Solche Nutzungen stellen keinen internen Gebrauch dar. In Frage kommen hier aber das *Zitatrecht* ([Art. 25 URG](#)) und die Schranke zugunsten der *Berichterstattung über aktuelle Ereignisse* ([Art. 28 Abs. 2 URG](#)). Beide Schranken setzen voraus, dass das Zitat bzw. der Ausschnitt aus Presseartikeln, Radio- oder Fernsehsendungen im Output bezeichnet wird, und die Quelle sowie gegebenenfalls die Urheberschaft angegeben werden ([Art. 25 Abs. 2](#) und [Art. 28 Abs. 2 URG](#)). Das ist zwar technisch anspruchsvoll, neuere Sprachmodelle sind aber durchaus in der Lage, diese formellen Voraussetzungen zu erfüllen. Materiell erfordert das Zitatrecht einen zulässigen Zitatzweck, das Zitat muss also zur Erläuterung, als Hinweis oder zur Veranschaulichung dienen ([Art. 25 Abs. 1 URG](#)). Zudem muss es in den Output des Systems eingebettet sein und ein inhaltlicher Bezug zwischen dem Output und dem zitierten Werk bestehen. Der Umfang des Zitats muss überdies durch den Zitatzweck gerechtfertigt sein.⁴² Die Schranke zugunsten der Berichterstattung über aktuelle Ereignisse erfordert in materieller Hinsicht nur, dass die Ausschnitte kurz sind und deren Nutzung der Information über aktuelle Fragen dient ([Art. 28 Abs. 2 URG](#)). Ob diese Voraussetzungen erfüllt sind, lässt sich nur im Einzelfall bestimmen.

IV. [RAG](#) mit Zugriff auf Internet

Neben der «klassischen» Nutzung von [RAG](#) mit Zugriff auf eine interne Datenbank wird der Ansatz auch genutzt, um auf Inhalte zuzugreifen, die im Internet abrufbar sind. Das Vorgehen kann dabei auf einer Identifikation der relevanten Inhalte beruhen, das demjenigen bei [RAG](#) mit Zugriff auf eine interne Datenbank weitgehend entspricht.⁴³ Bisweilen werden die relevanten Inhalte aber gar nicht spezifisch identifiziert, sondern einfach Teile von Webseiten (bspw. die ersten Absätze), die aufgrund einer Internetsuche (bspw. mit Google Search) als relevant identifiziert wurden, in einen Prompt übernommen.

Wie bei der Nutzung von [RAG](#) mit Zugriff auf eine interne Datenbank sind auch bei der Nutzung von [RAG](#) mit Zugriff auf das Internet aus urheberrechtlicher Sicht drei Schritte zu unterscheiden: (1) das Speichern der Werke auf einem Server und deren Zugänglichmachen, (2) der Zugriff auf diese Werke, die Identifikation der relevanten Inhalte und deren Integration in einen Prompt aufgrund einer Anfrage (*User query*) und (3) der Output des Sprachmodells. Beim Speichern von Werken auf einem Server und bei deren Zugänglichmachen im Internet stellen sich keine [RAG](#)-spezifischen Fragen. In der Folge wird deshalb nur skizziert, wie die Nutzung von Werken aufgrund einer Anfrage (*User query*) und im Output eines Sprachmodells urheberrechtlich zu beurteilen sind.

Dabei sind zwei Konstellationen zu unterscheiden. Zum einen können *Unternehmen, die [RAG](#) selber nutzen*, nicht nur auf Inhalte von internen Datenbanken zugreifen, sondern auch auf solche, die im Internet zugänglich sind. Diese Konstellation ist grundsätzlich gleich zu beurteilen wie die «klassische» Nutzung von [RAG](#); der einzige Unterschied besteht darin, dass sich die Frage nach der Zulässigkeit des Speicherns von Werken in einer internen Daten-

⁴² [BGE 131 III 480 ff. E. 2.1](#), «Schweizerzeit»; Thouvenin (Fn. 19), N 695; Rehbinders/Has/Huhlig (Fn. 16), [URG 25](#) N 6; Egloff/Künzi (Fn. 18), [URG 25](#) N 7; Hilty (Fn. 16), N 476.

⁴³ Siehe dazu vorne, II.

sic! 2025 S. 381, 389

bank nicht stellt.⁴⁴ Relevante Unterschiede ergeben sich aber, wenn Privatpersonen oder Unternehmen *Dienste Dritter* nutzen, bspw. ChatGPT, Gemini oder Copilot, und diese Dienste im Rahmen von [RAG](#) auf Inhalte zugreifen, die über das Internet abrufbar sind. Diese Frage ist hier näher zu untersuchen.

1. Zugriff auf Werke, Identifikation relevanter Inhalte und Integration in einen Prompt

Beim Zugriff auf Werke, die im Internet zugänglich sind, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt werden Vervielfältigungen erstellt. Da die Rechteinhaber diese Vervielfältigungen in aller Regel nicht erlaubt haben, stellt sich die Frage, ob sie von einer Schranke freigestellt werden. In Frage kommen die «Schranke»⁴⁵ für vorübergehende Vervielfältigungen ([Art. 24a URG](#)) und die Schranken zugunsten des Privatgebrauchs ([Art. 19 Abs. 1 lit. a URG](#)) sowie des internen Gebrauchs ([Art. 19 Abs. 1 lit. c URG](#)).

a) Vorübergehende Vervielfältigungen

Wie bereits für die Nutzung von [RAG](#) mit Zugriff auf eine interne Datenbank ausgeführt, sind die hier entstehenden Vervielfältigungen flüchtig und begleitend ([Art. 24a lit. a URG](#)) und sie stellen einen integralen und wesentlichen Teil eines technischen Verfahrens dar ([Art. 24a lit. b URG](#)).⁴⁶ Fraglich ist aber auch hier, ob die dritte und vierte Voraussetzung für die Qualifikation als vorübergehende Vervielfältigung erfüllt sind, ob die Vervielfältigungen also einer rechtmässigen Nutzung dienen ([Art. 24a lit. c URG](#)) und keine eigenständige wirtschaftliche Bedeutung ([Art. 24a lit. d URG](#)) haben.

Wie beim «klassischen» [RAG](#) dienen die Vervielfältigungen, die bei [RAG](#) mit Zugriff auf das Internet erstellt werden, nicht unmittelbar dem Werkgenuss, also der Wahrnehmung der Werke mit den menschlichen Sinnen. Eine *rechtmässige Nutzung* liegt damit nur vor, wenn eine Schranke greift. In Frage kommen dabei, je nach Nutzer des Dienstes, die Schranke zugunsten des Privatgebrauchs ([Art. 19 Abs. 1 lit. a URG](#)) oder des internen Gebrauchs ([Art. 19 Abs. 1 lit. c URG](#)).⁴⁷

Bei der Voraussetzung der *eigenständigen wirtschaftlichen Bedeutung* ist zu prüfen, ob Vervielfältigungen, die beim Zugriff auf Werke im Internet, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt erstellt werden, eine neue, eigenständige Nutzungsmöglichkeit eröffnen.⁴⁸ Das ist hier der Fall. Denn die Nutzung eines Chatbots, der auf Inhalte zugreift, die im Internet zugänglich sind, ermöglicht den Nutzern eines solchen Dienstes, auf eine Anfrage an den Chatbot eine weitgehend zuverlässige Antwort zu erhalten, statt die erforderlichen Informationen im Internet selbst zu suchen und auf deren Grundlage selbst eine Antwort zu erstellen. Die Nutzung eines solchen Chatbots vermag damit die Nutzung der im Internet zugänglichen Werke zu ersetzen. Soweit diese Werke nur gegen Bezahlung oder gegen den Konsum von Werbung genutzt werden können, verhindert deren Nutzung durch Chatbots die Monetarisierung dieser Werknutzungen durch die Rechteinhaber. Zudem erhöht die Nutzung von Werken mittels [RAG](#) die Zuverlässigkeit und Aktualität der Antworten von Chatbots und damit die Attraktivität dieser Angebote; [RAG](#) leistet damit einen wesentlichen Beitrag zum wirtschaftlichen Erfolg dieser Dienste. Die Nutzung hat damit eine eigenständige wirtschaftliche Bedeutung. Das lässt sich am Beispiel der *Medien* besonders gut veranschaulichen: Können Chatbots mithilfe von [RAG](#) auf die online zugänglichen Inhalte von Medien zugreifen, dann sind sie in der Lage, Anfragen von Nutzern tagesaktuell und mit hoher Zuverlässigkeit zu beantworten. Die Nutzer können sich so über das aktuelle Geschehen durch die Nutzung von Chatbots informieren und auf den direkten Konsum von Medien verzichten. Damit ist klar, dass die bei der Verwendung von [RAG](#) mit Zugriff auf das Internet erstellten Vervielfältigungen nicht als vorübergehende Vervielfältigungen im Sinn von [Art. 24a URG](#) qualifiziert werden können.

⁴⁴ Siehe dazu vorne, III.1

⁴⁵ Siehe dazu vorne, III.2.

⁴⁶ Siehe dazu vorne, III.2.a).

⁴⁷ Siehe dazu nachfolgend, IV.1.b).

⁴⁸ Siehe dazu vorne, III.2.a)dd).

b) Privatgebrauch oder interner Gebrauch

Je nachdem, ob die Dienste von Privatpersonen oder von Unternehmen genutzt werden, können die mit der Nutzung von [RAG](#) verbundenen Vervielfältigungen als Privatgebrauch ([Art. 19 Abs. 1 lit. a URG](#)) oder als interner Gebrauch ([Art. 19 Abs. 1 lit. c URG](#)) freigestellt sein. In beiden Konstellationen erfolgt der Eigengebrauch unter *Beizug des Anbieters des Dienstes als Dritter* ([Art. 19 Abs. 2 URG](#)).

Nach der Rechtsprechung des Bundesgerichts und der herrschenden (aber nicht unbestrittenen) Lehre greifen beim *Beizug von Dritten* stets die *Gegenausnahmen* ([Art. 19 Abs. 3 URG](#)), unabhängig davon, ob ein Privatgebrauch oder ein interner Gebrauch vorliegt.⁴⁹ Damit ist die vollständige oder weitgehend vollständige Vervielfältigung im Handel erhältlicher Werkexemplare an sich ausgeschlossen ([Art. 19 Abs. 3 lit. a URG](#)). Diese Auffassung greift allerdings zu kurz, weil es Konstellationen gibt, in denen es widersinnig und realitätsfremd wäre, die vollständige Vervielfältigung im Handel erhältlicher Werkexemplare wegen des Bezugs eines Dritten vom Eigengebrauch auszuschliessen, so namentlich beim Speichern ganzer Werke auf den Servern

sic! 2025 S. 381, 390

eines Cloud-Providers.⁵⁰ Richtigerweise wird man deshalb mit Blick auf den Dreistufentest zwischen Konstellationen differenzieren müssen, in denen Vervielfältigungen unter Beizug eines Dritten die normale Verwertung der Werke beeinträchtigen und solchen, in denen das nicht der Fall ist. Diese Streitfrage muss hier allerdings nicht vertieft werden. Denn bei den hier in Frage stehenden Vervielfältigungen führen das Verständnis der herrschenden Lehre und die differenzierende Auslegung mit Blick auf den Dreistufentest zum gleichen Ergebnis: Die *vollständige oder weitgehend vollständige Vervielfältigung im Handel erhältlicher Werkexemplare* durch KI-Systeme von Drittanbietern kann von den Schranken zugunsten des Privatgebrauchs und des internen Gebrauchs *nicht freigestellt* werden, weil dadurch die normale Verwertung der entsprechenden Werke beeinträchtigt würde. Wie sich bei der Prüfung der «Schranke» für vorübergehende Vervielfältigungen gezeigt hat, haben die Vervielfältigungen bei der Nutzung von [RAG](#) mit Zugriff auf das Internet eine eigenständige wirtschaftliche Bedeutung.⁵¹ Die engen inhaltlichen Bezüge zwischen diesem Tatbestandsmerkmal und der zweiten und dritten Stufe des Dreistufentests legen nahe, dass die normale Verwertung von Werken beeinträchtigt würde, wenn die Vervielfältigungen, die beim Zugriff auf Werke im Internet, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt, durch eine Schranke freigestellt würden. Besonders deutlich wird der Eingriff in die normale Verwertung auch hier, wenn man sich vor Augen führt, dass die Zulässigkeit der Nutzung von Medieninhalten durch die Anbieter von Chatbots dazu führen würde, dass diese Dienste die Angebote der Medien weitgehend substituieren könnten.

2. Werke im Output

Für die Frage nach der urheberrechtlichen Zulässigkeit der Wiedergabe von Werken oder Teilen von Werken im Output von KI-Systemen kann im Wesentlichen auf die Ausführungen zur Nutzung von [RAG](#) mit Zugriff auf eine interne Datenbank verwiesen werden.⁵² Namentlich werden auch bei der Nutzung von [RAG](#) mit Zugriff auf das Internet in der Regel weder Teile von Werken unverändert übernommen noch Werke generiert, die bestehenden Werken so ähnlich sind, dass sie in deren Schutzbereich fallen. Enthält der Output aber für sich geschützte Teile eines Werks oder ist er einem bestehenden Werk so ähnlich, dass er in dessen Schutzbereich fällt, und wurde er tatsächlich unter Verwendung eines bestehenden Werks generiert, stellt sich die Frage, ob die Nutzung durch eine *Schranke* freigestellt ist.

Auch für diese Frage kann auf die Ausführungen zur Nutzung von [RAG](#) mit Zugriff auf eine interne Datenbank verwiesen werden. In Frage kommen bei der Nutzung von KI-Systemen von Drittanbietern durch *Unternehmen* namentlich die Schranke zugunsten des internen Gebrauchs mitsamt den Gegenausnahmen

⁴⁹ [BGE 140 III 616 ff. E. 3.4.2 und E. 3.5.2](#), «Bibliothekslieferdienst»; Reh binder/Haas/Uhlig (Fn. 16), [URG 19](#) N 30 f.; Barrelet/Egloff (Fn. 17), [URG 19](#) N 21; Gasser (Fn. 18), [URG 19](#) N 28; Cherpillod (Fn. 32), N 790; Hilty (Fn. 16), N 464; a.M. M. Reh binder, Schweizerisches Urheberrecht, 3. Aufl., Bern 2000, 142; mit eingehender Kritik der h.L.: P.-E. Ruedin/X.-B. Ruedin, La copie de partitions à des fins personnelles, [sic! 2005, 795 ss](#); wie die h.L.: Botschaft URG 2006, BBI 2006, 3389 ff., 3429; ebenso schon Botschaft URG 1989, BBI 1989 III 477 ff., 541.

⁵⁰ Ebenso: EuGH vom 24. März 2022, C-433/20, «Austro-Mechano/Strato», N 34 ff; Thouvenin (Fn. 19), N 691; a.M.: C. Marti, Urheberrechtlicher Eigengebrauch unter Mitwirkung Dritter, Bern 2022, 188 ff.; Gasser (Fn. 18), [URG 19](#) N 25a.

⁵¹ Siehe dazu vorne, III.2.a).

⁵² Siehe dazu vorne, III.3.

([Art. 19 Abs. 1 lit. c i.V.m. Abs. 3 lit. a URG](#)), das Zitatrecht ([Art. 25 URG](#)) und die Schranke zugunsten der Berichterstattung über aktuelle Ereignisse ([Art. 28 Abs. 2 URG](#)).⁵³ Letztere steht im Vordergrund, wenn im Output eines KI-Systems Medieninhalte unverändert wiedergegeben werden. Soweit es sich dabei um kurze Ausschnitte handelt, die Wiedergabe der Information über aktuelle Ereignisse dient und die formellen Voraussetzungen erfüllt sind, ist die Wiedergabe von Medieninhalten im Output von KI-Systemen zulässig. Entscheidend ist mit Blick auf den Dreistufentest aber auch hier, dass die Schranke nicht so extensiv ausgelegt wird, dass die freigestellten Nutzungen die normale Verwertung von Medieninhalten beeinträchtigen.

Wenn *Privatpersonen* KI-Systeme nutzen, werden ihre Handlungen in aller Regel von der Schranke des Privatgebrauchs ([Art. 19 Abs. 1 lit. a URG](#)) erfasst, die jede Nutzung von Werken im Output (bspw. das Speichern auf einem eigenen Gerät) freistellt, sofern sie zu privaten Zwecken von einer Privatperson allein oder in einem Kreis eng verbundener Personen wie Verwandten oder Freunden erfolgt. Das gilt auch bei einer Speicherung des Outputs in der Cloud, die als Privatgebrauch unter Beizug des Cloud-Anbieters als Dritter ([Art. 19 Abs. 1 lit. a i.V.m. Art. 19 Abs. 2 URG](#)) zu qualifizieren ist. Eine Berufung auf das *Zitatrecht* oder die Schranke zugunsten der *Berichterstattung über aktuelle Ereignisse* ist nur erforderlich, wenn der Output den privaten Kreis verlässt oder zu nicht privaten Zwecken genutzt wird.

V. Erkenntnisse

Retrieval-Augmented Generation (RAG) eröffnet neue Möglichkeiten, um die Zuverlässigkeit und Aktualität des Outputs von Sprachmodellen zu erhöhen. *RAG* ist bereits weit verbreitet, wurde aber noch kaum urheberrechtlich untersucht. Bei der urheberrechtlichen Beurteilung sind zwei Konstellationen zu unterscheiden: *RAG* mit Zugriff auf eine interne Datenbank und *RAG* mit Zugriff auf das Internet.

Bei *RAG mit Zugriff auf eine interne Datenbank* stellen sich drei Fragen: (1) Beim *Speichern von Werken in der Datenbank* ergeben sich keine Besonderheiten. Die damit verbundenen Vervielfältigungen sind zulässig, wenn die Rechteinhaber zugestimmt haben oder die Werke im Rahmen der Schranke zugunsten des internen Gebrauchs, also für die interne Information oder Dokumentation, verwendet werden ([Art. 19 Abs. 1 lit. c URG](#)). Nicht freigestellt ist dabei die vollständige oder weitgehend vollständige Vervielfältigung im Handel erhältlicher Werkexemplare ([Art. 19 Abs. 3 lit. a URG](#)). (2) Beim *Zugriff auf die Werke, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt* werden ebenfalls Vervielfältigungen erstellt. Diese können nicht allgemein als vorübergehende Vervielfältigungen im Sinn von [Art. 24a URG](#) qualifiziert werden, weil sie nur bisweilen einer rechtmässigen Nutzung dienen und

sic! 2025 S. 381, 391

ihnen wohl meist eine eigenständige wirtschaftliche Bedeutung zukommt. Die Vervielfältigungen sind dennoch zulässig, sofern ein interner Gebrauch vorliegt und keine im Handel erhältlichen Werkexemplare vollständig oder weitgehend vollständig vervielfältigt werden ([Art. 19 Abs. 1 lit. c i.V.m. Art. 19 Abs. 3 lit. a URG](#)). (3) Der *Output von Sprachmodellen* kann Teile von Werken enthalten oder bestehenden Werken so ähnlich sein, dass er in deren Schutzbereich fällt. In der Regel wird dies allerdings nicht der Fall sein, weil Sprachmodelle keine Texte duplizieren, sondern neu erstellen. Fällt der Output ausnahmsweise in den Schutzbereich eines bestehenden Werks und wurde dieses bei der Erstellung des Outputs tatsächlich genutzt, liegt zwar eine Urheberrechtsverletzung vor. Die Nutzung kann aber in vielen Fällen als interner Gebrauch ([Art. 19 Abs. 1 lit. c URG](#)), als Zitat ([Art. 25 URG](#)) oder als Berichterstattung über aktuelle Ereignisse ([Art. 28 Abs. 2 URG](#)) freigestellt werden. Die Nutzung von Werken Dritter im Rahmen von *RAG* mit Zugriff auf eine interne Datenbank ist damit zulässig, wenn sie für den internen Gebrauch, also für die interne Information oder Dokumentation, erfolgt, und keine im Handel erhältlichen Werkexemplare vollständig oder weitgehend vollständig vervielfältigt werden. Andernfalls muss die Zustimmung der Rechteinhaber eingeholt werden.

Bei *RAG mit Zugriff auf das Internet* werden keine Werke in internen Datenbanken gespeichert. Für die Nutzer und Anbieter eines solchen Dienstes stellen sich damit nur zwei Fragen: (1) Wie bei *RAG* mit Zugriff auf eine interne Datenbank werden auch hier beim *Zugriff auf Werke, bei der Identifikation der relevanten Inhalte und bei deren Integration in einen Prompt* Vervielfältigungen erstellt. Wie beim «klassischen» *RAG* lassen sich diese auch hier nicht als vorübergehende Vervielfältigungen im Sinn von [Art. 24a URG](#) qualifizieren. Nutzen Privatpersonen Angebote Dritter, bspw. einen Chatbot wie ChatGPT, können die im Rahmen von *RAG* erstellten Vervielfältigungen zwar als Privatgebrauch freigestellt sein ([Art. 19 Abs. 1 lit. a URG](#)). Nicht freigestellt ist wegen des Beizugs eines Dritten aber auch hier die vollständige oder weitgehend vollständige Vervielfältigung im Handel erhältlicher Werkexemplare ([Art. 19 Abs. 2 i.V.m. Art. 19 Abs. 3 lit. a](#)

⁵³ Siehe dazu vorne, III.3.

URG). Dasselbe gilt für die Nutzung solcher Dienste durch Unternehmen im Rahmen des internen Gebrauchs ([Art. 19 Abs. 1 lit. c i.V.m. Art. 19 Abs. 3 lit. a URG](#)). (3) Für den *Output* gilt grundsätzlich dasselbe wie beim «klassischen» [RAG](#). Je nachdem, ob der Dienst durch Privatpersonen oder Unternehmen genutzt wird, greift die Schranke des Privatgebrauchs oder des internen Gebrauchs, wenn der Output ausnahmsweise in den Schutzbereich eines bestehenden Werks fällt und dieses beim Erstellen des Outputs tatsächlich genutzt worden ist. Auch hier kann die Nutzung zudem als Zitat ([Art. 25 URG](#)) oder als Berichterstattung über aktuelle Ereignisse ([Art. 28 Abs. 2 URG](#)) freigestellt sein. Die Nutzung von Werken Dritter im Rahmen von [RAG](#) mit Zugriff auf das Internet ist damit zulässig, wenn die Vervielfältigungen vom Anbieter eines Dienstes im Rahmen des Privatgebrauchs oder des internen Gebrauchs erstellt und keine im Handel erhältlichen Werkexemplare vollständig oder weitgehend vollständig vervielfältigt werden. Andernfalls muss auch hier die Zustimmung der Rechteinhaber eingeholt werden.

sic! 2025 S. 381, 392

Zusammenfassung

Retrieval-Augmented Generation ([RAG](#)) eröffnet neue Möglichkeiten, um die Zuverlässigkeit und Aktualität des Outputs von Sprachmodellen zu erhöhen. [RAG](#) ist bereits weit verbreitet, wurde aber noch kaum urheberrechtlich untersucht. Bei der urheberrechtlichen Beurteilung sind zwei Konstellationen zu unterscheiden: [RAG](#) mit Zugriff auf eine interne Datenbank und [RAG](#) mit Zugriff auf das Internet.

Bei [RAG mit Zugriff auf eine interne Datenbank](#) greifen Unternehmen beim Einsatz eines Sprachmodells auf Dokumente zu, die sie in einer Datenbank gespeichert haben. Soweit es sich dabei um Werke handelt, an denen Dritte urheberrechtlich berechtigt sind, stellt sich die Frage, ob das Speichern der Dokumente und deren Nutzung von einer Schranke freigestellt sind. Beim Einsatz von [RAG](#) werden Vervielfältigungen erstellt. Diese sind zwar flüchtig und technisch bedingt, können aber nicht allgemein als vorübergehende Vervielfältigungen im Sinn von [Art. 24a URG](#) qualifiziert werden, weil sie nur bisweilen einer rechtmässigen Nutzung dienen und der Nutzung im Rahmen von [RAG](#) wohl meist eine eigenständige wirtschaftliche Bedeutung zukommt. Zulässig sind die Vervielfältigungen aber, wenn sie im Rahmen des internen Gebrauchs erfolgen, wenn die Nutzung von [RAG](#) also der Information oder Dokumentation einer Organisation dient und keine im Handel erhältlichen Werkexemplare vollständig oder weitgehend vollständig vervielfältigt werden ([Art. 19 Abs. 1 lit. c i.V.m. Abs. 3 lit. a URG](#)).

Bei [RAG mit Zugriff auf das Internet](#) nutzen Privatpersonen oder Unternehmen einen Dienst, bspw. ChatGPT, Gemini oder Copilot, der auf Inhalte zugreift, die über das Internet abrufbar sind. Wie bei [RAG mit Zugriff auf eine interne Datenbank](#) können die dabei entstehenden Vervielfältigungen nicht allgemein als vorübergehende Vervielfältigungen im Sinn von [Art. 24a URG](#) qualifiziert werden. Zulässig sind die Vervielfältigungen aber, wenn sie im Rahmen des Privatgebrauchs oder des internen Gebrauchs unter Beizug des Diensteanbieters als Dritter erstellt ([Art. 19 Abs. 1 lit. a bzw. lit. c i.V.m. Abs. 2 URG](#)) und keine im Handel erhältlichen Werkexemplare vollständig oder weitgehend vollständig vervielfältigt werden ([Art. 19 Abs. 2 i.V.m. Abs. 3 lit. a URG](#)).

Der Output von Sprachmodellen kann Teile von Werken enthalten oder bestehenden Werken so ähnlich sein, dass er in deren Schutzbereich fällt. In der Regel wird dies zwar nicht der Fall sein, weil Sprachmodelle nicht Texte duplizieren, sondern neu erstellen. Das gilt für Sprachmodelle allgemein und auch beim Einsatz von [RAG](#). Fällt der Output aber in den Schutzbereich eines bestehenden Werks und wurde dieses bei dessen Erstellung tatsächlich genutzt, liegt grundsätzlich eine Urheberrechtsverletzung vor. Die Nutzung kann aber als Privatgebrauch ([Art. 19 Abs. 1 lit. a URG](#)), als interner Gebrauch ([Art. 19 Abs. 1 lit. c URG](#)), als Zitat ([Art. 25 URG](#)) oder als Berichterstattung über aktuelle Ereignisse ([Art. 28 Abs. 2 URG](#)) freigestellt sein, wenn die jeweiligen Voraussetzungen erfüllt sind.&cbr;

Résumé

La génération augmentée de récupération (RAG) ouvre de nouvelles possibilités pour augmenter la fiabilité et l'actualité des réponses fournies par les modèles de langage. La RAG est déjà largement répandue, mais elle n'a guère été étudiée sous l'angle du droit d'auteur. Lors de l'évaluation sous l'angle du droit d'auteur, il convient de distinguer deux constellations: RAG avec accès à une base de données interne et RAG avec accès à Internet.

Dans le cas de la RAG ayant accès à une base de données interne, les entreprises accèdent à des documents qu'elles ont stockés dans une base de données lorsqu'elles utilisent un modèle de langage. Dans la mesure où il s'agit d'œuvres pour lesquelles des tiers sont titulaires de droits d'auteur, la question se pose de savoir si le stockage des documents et leur utilisation sont exemptés par une restriction. Lors de l'utilisation de la RAG, des reproductions sont réalisées. Celles-ci sont certes fugaces et conditionnées par la technique, mais ne peuvent pas être qualifiées de manière générale de reproductions provisoires au sens de l'[art. 24a LDA](#), car elles ne servent qu'occasionnellement à une utilisation licite et l'utilisation dans le cadre de la RAG revêt le plus souvent une importance économique propre. Les reproductions sont toutefois

autorisées lorsqu'elles sont effectuées dans le cadre d'un usage interne, c'est-à-dire lorsque l'utilisation de la RAG sert à l'information et à la documentation et qu'aucun exemplaire d'œuvre disponible dans le commerce n'est reproduit intégralement ou en grande partie (art. 19 al. 1 let. c en relation avec l'al. 3 let. a [LDA](#)).

Dans le cas de la RAG avec accès à Internet, les particuliers ou les entreprises utilisent les services de tiers, par exemple ChatGPT, Gemini ou Copilot, qui, dans le cadre de la RAG, accèdent à des contenus disponibles sur Internet. Comme dans le cas de la RAG avec accès à une base de données interne, les reproductions qui en résultent ne peuvent pas être qualifiées de manière générale de reproductions provisoires au sens de l'[art. 24a LDA](#). Les reproductions sont toutefois autorisées si elles sont réalisées dans le cadre d'une utilisation privée ou interne avec le concours du fournisseur de services en tant que tiers (art. 19 al. 1 let. a et let. c en relation avec l'al. 2 [LDA](#)) et si aucun exemplaire d'œuvre disponible dans le commerce n'est reproduit intégralement ou en grande partie (art. 19 al. 2 en relation avec l'al. 3 let. a [LDA](#)).

Les réponses fournies par les modèles de langage peuvent contenir des parties d'œuvres ou être si similaires à des œuvres existantes qu'elles tombent dans leur domaine de protection. En règle générale, ce ne sera pas le cas, car les modèles de langage ne dupliquent pas des textes, mais en créent de nouveaux. Cela vaut pour les modèles de langage en général et également pour l'utilisation de la RAG. Mais si les réponses fournies tombent dans le domaine de protection d'une œuvre existante et que celle-ci a effectivement été utilisée lors de sa création, il y a en principe violation du droit d'auteur. L'utilisation peut toutefois être exemptée en tant qu'utilisation à des fins privées ([art. 19 al. 1 let. a LDA](#)), en tant qu'utilisation interne ([art. 19 al. 1 let. c LDA](#)), en tant que citation ([art. 25 LDA](#)) ou en tant que compte rendu d'actualité ([art. 28 al. 2 LDA](#)), si les conditions requises sont remplies.