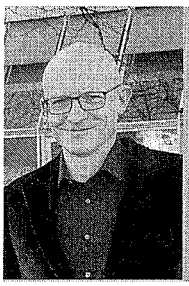
LegalTech

Zeitschrift für die digitale Rechtsanwendung [LTZ]
3/2024 | Seiten 235-239

Können große Sprachmodelle Jura?

Ministerialrat Dr. Sebastian Dötterl

Das Potential von Large Language Models (LLMs) im Rechtswesen wird als besonders hoch eingeschätzt. Meldungen über vermeintliche Top-Ergebnisse im Bar Exam einerseits und angeblich massenhaftes Halluzinieren bei rechtlichen Fragestellungen andererseits ziehen immer wieder Aufmerksamkeit auf sich. Dieser Beitrag stellt wichtige Studien zum Einsatz von LLMs im Rechtsbereich vor, versucht sie einzuordnen und zieht Schlussfolgerungen zum Potential und zum künftigen Einsatz von LLMs.



Der Autor ist Referatsleiter im Bayerischen Staatsministerium der Justiz. Er befasst sich unter anderem mit Fragen rund um Digitalisierung und Künstliche Intelligenz, insbesondere im Kontext von Justiz und Rechtsstaat. Der Beitrag gibt seine persönliche Meinung wieder.

I. GPT-4 und das vermeintliche Top-Ergebnis im Bar Exam

Im März 2023 war zu lesen, GPT-4 übertreffe bei den Prüfungen zum US Bar Exam 90 % der menschlichen Prüflinge.[1] Ursprung dieser Aussage dürfte eine Studie sein, in der die Leistung des Modells GPT-4 in allen Teilen der Uniform Bar Examination (UBE) untersucht wurde.[2]

1. GPT-4 besteht das Bar Exam...

a) Aufgabe

Das UBE besteht aus

- einer sechsstündigen Multiple-Choice-Prüfung mit 200 Fragen (Multistate Bar Examination, MBE),
- einem dreistündigen Essay-Teil, bei dem sechs Essays verfasst werden müssen (Multistate Essay Exam, MEE) und
- einem dreistündigen Teil mit zwei Fallstudien mit fiktivem Aktenmaterial und Rechtsvorschriften (Multistate Performance Test, MPT).[3]

b) Vorgehensweise

GPT-4 wurde per *zero-shot*-Prompting (also ohne weitere Hilfestellungen wie eine Rollenzuweisung oder Beispiele) mit den Aufgaben konfrontiert. Die einzige Erleichterung war, dass einzelne Unterfragen etwa zu den Essays aufgeteilt und einzeln geprompted wurden.

Die Prüfungsleistungen von GPT-4 wurden von den rechtlich qualifizierten Studienautoren bewertet. Zusätzlich wurden Blindvergleiche von anderen Experten eingeholt.

c) Ergebnis

GPT-4 überschritt mit der erreichten Punktzahl von ca. 297 Punkten die Bestehensgrenze aller UBE-Jurisdiktionen, die meist zwischen 260 und 280 Punkten liegt, signifikant. Darin liegt eine deutliche Verbesserung zum Vorgängermodell GPT-3.5, das mit nur 213 Punkten klar scheitern würde.[4]

2. ... aber nicht in den Top 10 %

a) Herkunft der Behauptung

Die Studie „GPT passes the Bar Exam" hatte nicht die primäre Zielsetzung, die Leistung von GPT-4 mit dem Ergebnis der menschlichen Prüflinge zu vergleichen. Im Preprint der Studie findet sich insoweit nur eine Nebenbemerkung in einer Fußnote: "*Using a percentile chart from a recent exam administra-*

1 Koetsier, GPT-4 Beats 90 % Of Lawyers Trying To Pass The Bar, https://www.forbes.com/sites/johnkoetsier/2023/03/14/gpt-4-beats-90-of-lawyers-trying-to-pass-the-bar/ (letzter Zugriff auf alle Internetquellen am 17.5.2024).

2 Katz/Bommarito/Gao/Arredondo, GPT-4 passes the bar exam, Phil. Trans. R. Soc. A.382, https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0254.

3 vgl. https://www.ncbex.org/exams/ube

4 Katz/Bommarito/Gao/Arredondo, GPT-4 passes the bar exam, Phil. Trans. R. Soc. A.382, S. 12, https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0254..

tion [...], GPT-4 would receive a combined score approaching the 90th percentile of test-takers."[5]

Dies wurde dann von OpenAI offensiv vermarktet[6] und in den Medien aufgegriffen.[7]

b) Kritik in späterer Studie

Die Bewertung der GPT-4 Performance im Bar Exam wurde in einer weiteren Studie methodisch hinterfragt.[8]

Ein Kritikpunkt betraf die Bewertung der Essays und Fallstudien in der ersten Studie. Diese habe nicht den Korrekturstandards entsprochen, die im UBE angewendet werden. Der Schwerpunkt der Methodenkritik richtete sich aber gegen die Einordnung von GPT-4 in die 90. Perzentile bzw. Top-10 % der Absolventen. Diese Einordnung sei nicht anhand einer repräsentativen Vergleichsgruppe von Absolventen errechnet worden, sondern anhand eines Datensatzes, der zu einem großen Teil Wiederholer enthielt, die im ersten Anlauf gescheitert waren. Mit anderen Worten: GPT-4 sei mit einer leistungsschwachen Gruppe verglichen worden, womit seine Leistung überbewertet wurde.

Der Autor des Papers stellte eigene Berechnungen anhand ausgewogenerer Daten an und kam zum Ergebnis, dass das Ergebnis von GPT-4 insgesamt die 68. Perzentile erreicht habe.[9] Bei einer weiteren Einengung der Vergleichsgruppe auf diejenigen, die die Prüfung tatsächlich bestanden hatten, fiel die Leistung von GPT-4 in etwa in den Medianbereich.[10]

Differenziere man nun noch zwischen Multiple-Choice-Teil und den anderen beiden Prüfungsteilen, liege das Ergebnis von GPT-4 bei Multiple-Choice in der 69. Perzentile, bei den beiden Teilen mit „offenen" Aufgaben jedoch nur etwa in der 15. Perzentile.

3. Bewertung

GPT-4 erzielte im Bar Exam Leistungen, die deutlich zum Bestehen reichen dürften. Dabei zeigte GPT-4 seine Stärken im Multiple-Choice-Teil und war bei den „offenen" Fragestellungen eher schwach. Insgesamt stellte GPT-4 gegenüber GPT-3.5 einen Quantensprung dar. Das von OpenAI im Anschluss reklamierte Spitzenergebnis ist jedoch übertrieben.

II. LLMs und Halluzinationen

Laut einer im Januar 2024 als Preprint veröffentlichten Studie von Forschenden der Stanford University halluzinieren moderne Sprachmodelle in zwischen 69 und 88 % der Fälle, wenn sie Fragen zu Fällen aus dem Spektrum der US-Bundesgerichte beantworten.[11]

1. Große Sprachmodelle halluzinieren in bis zu 88 % der Fälle...

a) Aufgabe

Die untersuchten LLMs ChatGPT 3.5, LLama 2 und PaLM 2 wurden mit einer Vielzahl von Fragen zu Gerichtsfällen gepromptet. Die Aufgaben waren in drei Komplexitätsstufen eingeteilt.

- Auf der niedrigsten Stufe sollten die Modelle zB anhand eines gegebenen Fallnamens und der Fundstelle bestimmen, ob der Fall wirklich existiert, welches Gericht geurteilt hatte und wer Berichterstatter war.
- Auf der mittleren Stufe wurden inhaltliche Fragen zu den Fällen gestellt, wie zB ob das Gericht die Entscheidung der Vorinstanz bestätigte oder aufhob und welche Präzedenzfälle zitiert wurden.
- Auf der höchsten Komplexitätsstufe sollten die Modelle zB anhand zweier gegebener Fallnamen und Fundstellen beurteilen, ob die rechtliche Bewertung der Fälle übereinstimmte oder sich widersprach, und Fragen zum Sachverhalt, zur Verfahrensgeschichte und zur zentralen Rechtsfrage beantworten.

b) Vorgehensweise

Die Modelle wurden sowohl mit *zero-shot* als auch mit *three-shot*-Prompts gefüttert.[12] Um die Richtigkeit der Antworten zu bewerten, nutzten die Forschenden je nach Komplexität der Aufgabenstellung zwei Herangehensweisen:

- Bei den einfacheren Aufgaben wurden die Antworten mit einem für die Untersuchung aufbereiteten Datensatz verglichen (zB das urteilende Gericht, zitierte Präzedenzfälle, etc).
- Bei schwierigeren Aufgaben, für die keine Metadaten vorlagen, wurde das Modell zweimal mit der gleichen Aufgabe konfrontiert. Anschließend setzten die Forschenden das Modell GPT-4 (!) ein, um zu bewerten, ob sich die beiden Antworten widersprachen. Zeigte sich ein Widerspruch, wurde dies als Halluzination bewertet, da zwei sich widersprechende Antworten nicht beide richtig sein können.[13]

c) Ergebnis

Je nach Komplexität, Aktualität und Spruchkörper unterschieden sich die Ergebnisse. Je komplexer die Aufgabe, desto mehr Halluzinationen wurden festgestellt. Bei neueren Fällen waren

5 Katz/Bommarito/Gao/Arredondo, GPT-4 passes the bar exam, Phil. Trans. R. Soc. A.382, https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0254; Preprint abrufbar unter https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389233 dort S. 10, Fn. 3; Diese Aussage wurde in der Endfassung der Studie abgeschwächt:*"within a range between 68th and 90th percentile (depending on the precise state and timing of the exam administration)"*.

6 OpenAI u.a.: GPT-4 Technical Report, https://arxiv.org/abs/2303.08774, S. 30; ferner zB https://openai.com/index/gpt-4-research/: "[GPT-4] passes a simulated bar exam with a score around the top 10 % of test takers".

7 Vgl. Koetsier, GPT-4 Beats 90 % Of Lawyers Trying To Pass The Bar, https://www.forbes.com/sites/johnkoetsier/2023/03/14/gpt-4-beats-90-of-lawyers-trying-to-pass-the-bar/.

8 Martínez, E. Re-evaluating GPT-4's bar exam performance Artif Intell Law (2024), https://doi.org/10.1007/s10506-024-09396-9.

9 Insoweit haben die Autoren der ersten Studie die im Preprint noch enthaltene Bewertung abgeschwächt, vgl. Fn. 5.

10 Hier nennt Martínez einerseits im Abstract und im Fazit die 48. Perzentile, in der eigentlichen Arbeit (S. 11, unter 3.2.2.) andererseits die 45. Perzentile, Martínez, E. Re-evaluating GPT-4's bar exam performance Artif Intell Law (2024).

11 Dahl/Magesh/Suzgun/Ho, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, https://arxiv.org/abs/2401.01301.

12 Three-Shot: Das Modell erhält zusätzlich zur Aufgabenstellung drei Beispiele, die die gewünschte Antwortweise veranschaulichen.

13 Wie die Studie klarstellt, werden durch diese Vorgehensweise konsistent falsche Antworten nicht als solche entlarvt, vgl. Dahl/Magesh/Suzgun/Ho, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, S. 8, https://arxiv.org/abs/2401.01301. Diese erlaubt daher nur die Feststellung einer Untergrenze für die tatsächliche Halluzinationsrate.

die Antworten tendenziell besser als bei alten Fällen, bei Fällen des Supreme Courts wurden bessere Ergebnisse erzielt als bei niedrigeren Instanzen.

Insgesamt zeigte sich, dass alle untersuchten Modelle häufig falsche Antworten gaben. Betrachtet man die Aufgabenstellungen, bei denen die Antwort anhand der Metadaten überprüft wurde (vgl. oben II. 2 lit. b), ergab sich für GPT-3.5 eine Halluzinationsrate von 69 %, für PaLM 2 von 72 % und für Llama 2 von 88 %. Zudem präsentierten die Modelle ihre teils falschen Antworten oft mit großem Selbstbewusstsein.[14]

2. ... wenn veraltete Modelle in ungeeigneter Weise benutzt werden

Die Studie verwendete zum Zeitpunkt des Erscheinens bereits veraltete Sprachmodelle wie zB GPT-3.5.[15]

Zudem wurden die Modelle mit einer Aufgabenstellung konfrontiert, mit der sie bekanntermaßen ohne weitere Hilfe wie etwa Zugriff auf eine Datenbank nicht gut zurechtkommen, nämlich mit der Abfrage konkreter Wissensdetails. Letztlich wurden die LLMs hier als eine Art Suchmaschine oder Datenbank genutzt.

3. Bewertung

Die hohen Halluzinationsraten im gewählten Studiendesign unterstreichen, dass Sprachmodelle ohne weitere Maßnahmen für derartige Wissensabfragen ungeeignet sind.

III. LLMs und Kostenersparnis

In einem im Januar 2024 veröffentlichten Paper wurde untersucht, wie leistungsfähig LLMs bei der Überprüfung von Verträgen im Vergleich zu menschlichen Juristen sind.[16] Das Paper kommt zum Ergebnis, dass LLMs qualitativ auf Augenhöhe mit *junior lawyers* agierten und das zu einem Bruchteil der Kosten.

1. LLMs sparen LLMs 99,97 % der Kosten ein...

a) Aufgabe

Die LLMs wurden mit der Analyse von zehn echten, anonymisierten Beschaffungsverträgen beauftragt, die sowohl aus den USA als auch aus Neuseeland stammten. Die Aufgabe der LLMs bestand darin, festzustellen, ob der Vertrag den vordefinierten Standards entsprach oder davon abwich, wobei das LLM entweder die Perspektive des Käufers oder des Lieferanten einnehmen sollte.

b) Vorgehensweise

Die LLMs, darunter GPT-4, wurden in den Prompts mit umfassenden Informationen ausgestattet. Dazu gehörten

- die Vertragstexte,
- Angaben zur anzuwendenden Rechtsordnung,
- Informationen zu den beteiligten Unternehmen (Größe, Branche, Art des Produkts) und
- ein *contract review playbook* mit standardisierten Prüfkriterien.

Die Leistung der LLMs wurde mit der der *junior lawyers* verglichen, und zwar hinsichtlich Genauigkeit, Geschwindigkeit und Kosten. Als Vergleichsmaßstab wurden die Bewertungen erfahrener Juristinnen (*senior lawyers*) herangezogen.

c) Ergebnis

Die Veröffentlichung kam zu dem Ergebnis, dass die LLMs qualitativ auf Augenhöhe mit *junior lawyers* abschnitten.

Gleichzeitig benötigten die LLMs nur einen Bruchteil der Zeit für die Aufgabe. Während Menschen eine Stunde oder länger benötigten, schlossen die schnellsten LLMs die Prüfung in unter einer Minute ab.

Daraus ergab sich – wenig überraschend – eine enorme Kostenersparnis. Das Paper errechnet eine Reduktion von über 99,9 % gegenüber der Befassung von *junior lawyers*.

2. ...jedoch bleibt die Methodik der Untersuchung unklar

Allerdings ist die Methodik anhand der veröffentlichten Informationen nicht gut nachvollziehbar. Es fehlen Angaben zur Stichprobengröße und -auswahl, Bewertungskriterien, Inhalt des *contract review playbook*, Qualifikation und Anzahl der *junior lawyers* und einiges mehr. Auch sollte im Hinterkopf behalten werden, dass der Preprint von Mitarbeitenden der Firma Onit verfasst wurde, einer Firma, die selbst Legal Tech-Produkte herstellt.[17]

Die errechnete Kostenersparnis könnte daher übertrieben sein.

3. Bewertung

Das Paper zeigt dennoch großes Potential von LLMs bei der Vertragsprüfung, wenn sie gut „bepromptet" und mit zusätzlichen Informationen und Kontext versorgt werden.

IV. LLMs und Studienprüfungen

In einer im November 2023 online veröffentlichten Studie von Forschenden der University of Minnesota Law School wurde der Effekt von KI-Unterstützung bei Studierenden untersucht.[18] Es ergab sich je nach Aufgabentyp eine Zeitersparnis um bis zu 32 %.

1. GPT-4 macht Studierende schneller...

a) Aufgabe

60 Studierende wurden zufällig in zwei Gruppen von je 30 Personen eingeteilt. Die Teilnehmenden musste vier Aufgaben bearbeiten: Verfassen einer Klageschrift, eines Vertrags, eines Abschnitts eines Mitarbeiterhandbuchs und eines Mandantenschreibens.

14 Dahl/Magesh/Suzgun/Ho, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, S. 13 f., https://arxiv.org/abs/2401.01301.

15 Hingegen wurde zur Bewertung von Antworten GPT-4 genutzt, s. oben II. 1. lit. b), zweiter Spiegelstrich.

16 Martin/Whitehouse/Yiu/Catterson/Perera, Better Call GPT, Comparing Large Language Models Against Lawyers, https://arxiv.org/pdf/2401.16212.pdf.

17 vgl. www.onit.com

18 Choi/Monahan/Schwarcz, Lawyering in the Age of Artificial Intelligence, 109 Minnesota Law Review (Veröffentlichung für 2024 angekündigt), https://ssrn.com/abstract=4626276.

b) Vorgehensweise

Jede Gruppe hatte im Wechsel GPT-4 als Unterstützung zur Verfügung, während die jeweils andere Gruppe klassisch arbeitete. Vorher hatten die Teilnehmenden ein mehrstündiges Training zur effektiven Nutzung von GPT-4 absolviert. Die Prüfungsarbeiten wurden blind bewertet.

c) Ergebnis

Der Einsatz von GPT-4 führte bei allen vier Aufgaben zu einer deutlichen Verringerung der Bearbeitungszeit zwischen 12 % und 32 %. Die Teilnehmenden berichteten zudem eine erhöhte Zufriedenheit bei der Aufgabenbearbeitung mit KI.

2. ...aber nicht unbedingt besser

Der Einsatz von GPT-4 erhöhte laut Studie die Qualität der abgegebenen Arbeiten nur geringfügig und ungleichmäßig. Wo es Qualitätsverbesserungen gab, waren diese ungleich verteilt: Leistungsschwächere Teilnehmende profitierten mehr als leistungsstärkere. KI war – jedenfalls im vorliegenden Studiensetting – eine Art Gleichmacher.[19]

3. Bewertung

Die Studie zeigt, dass LLMs das Potential haben, die Produktivität bei juristischer Arbeit zu steigern. Mithilfe von KI-Assistenz können in kürzerer Zeit qualitativ gleichwertige Arbeitsergebnisse erzielt werden.

V. Retrieval Augmented Generation (RAG) löst nicht alle Probleme.

In einem Preprint aus dem Mai 2024 wurden juristische Recherchetools untersucht, die auf RAG, also den Zugriff der Sprachmodelle auf qualitativ hochwertige juristische Daten, setzen. Solche Angebote von LexisNexis und Thomson Reuters sind bereits auf dem Markt. Dabei ergaben sich bei 17-33% der Anfragen falsche Antworten.[20]

Die Genauigkeit und Verlässlichkeit bleibt also auch bei RAG ein Thema.

VI. Schlussfolgerungen

1. Eine Momentaufnahme

Die vorgestellten Studien stellen angesichts der rasanten technologischen Entwicklung nur eine Momentaufnahme dar, teilweise auch nur einen historischen Rückblick.

Sämtliche Untersuchungen basieren auf englischsprachigen Aufgabenstellungen und Rechtssystemen. Die Ergebnisse, insbesondere die festgestellten Effizienzgewinne, lassen sich nicht ohne Weiteres auf die deutsche Sprache und das deutsche Recht übertragen, da weniger deutschsprachiges Trainingsmaterial zur Verfügung steht.

Ein strukturiertes Subsumieren von Normen in deutscher Sprache anhand eines konkreten Falls stellt die aktuellen LLMs nach den Feldversuchen des Verfassers noch vor große Probleme. Zu den Ergebnissen eines von Dr. Ann-Kristin Mayrhofer und dem Verfasser gemeinsam veranstalteten Seminars an der Ludwig-Maximilians-Universität München, bei dem die Studierenden für ihre deutschsprachigen Seminararbeiten KI-Tools nutzen und den Einsatz dokumentierten, siehe das *Editorial* dieser Ausgabe.

2. KI als Game-Changer? Es kommt darauf an...

KI ist kein Selbstzweck, sondern ihr Einsatz muss gut überlegt sein. Für einen sinnvollen Einsatz kommt es auf eine Reihe von Faktoren an, zB:[21]

- Use-Case,
- konkret verwendetes Sprachmodell,
- Pre-training und/ oder Finetuning mit rechtsspezifischem Material[22],
- Prompting-Technik,
- Zugriff des Modells auf externe, kuratierte Daten.[23]

3. Offene Fragen

Offen ist, inwieweit die Leistungsfähigkeit der LLMs sich weiter steigern lässt (auch in Kombination mit regelbasierten Ansätzen, sog. „hybride KI"[24]), oder an welchem Punkt sie an eine Grenze stoßen. Für die juristische Praxis relevant ist dabei insbesondere, inwieweit sich die Probleme um Halluzinationen, mangelnde Nachvollziehbarkeit (*black box*) und Verzerrungen in den Trainingsdaten (*bias*) lösen oder abmildern lassen. Relevant ist ferner, inwieweit das Erfassen komplexerer Sachverhalte (etwa mit mehreren Personen und Gegenständen), das Erkennen irrelevanter oder überholter Informationen (z.B. aufgehobene Entscheidungen, veraltetes Recht) sowie die Subsumtion im Gutachtenstil gelingen können.

Offen ist ferner,

- wie sich die Kosten des Einsatzes entwickeln (Energiekosten, Softwarelizenzen, Personalkosten für KI-Training, -Betreuung, -Schulungen, Compliance, Haftungsrisiken),
- welche Auswirkungen die noch ungeklärten urheberrechtlichen[25] und datenschutzrechtlichen[26] Grundsatzfragen rund um das LLM-Training in der Praxis haben werden und
- wie sich die gesellschaftlichen Einstellungen zu KI entwickeln werden – allgemein, und spezifisch im Rechtswesen.

19 Nach einer Veröffentlichung von EY nutzen Studierende, die ihre eigenen Leistungen als "überdurchschnittlich" oder "exzellent" einschätzen, im Rahmen ihres Studiums mit höherer Wahrscheinlichkeit KI-Tools als "durchschnittliche" oder "unterdurchschnittliche" Studierende, vgl. Pressemitteilung vom 26.3.2024, https://www.ey.com/de_de/news/2024/03/ey-studierendenstudie-2024-kuenstliche-intelligenz.

20 Magesh et al: Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf, sowie allgemein Schomerus/Leenen, unternehmensjurist 2024, 32 (34), Burkert/Bachmann/Schomerus/Leenen/Walinski, unternehmensjurist 2024, 36 (39).

21 Vgl. auch Schomerus/Leenen unternehmensjurist 2024, 32 und speziell zum Steuerbereich Braun/Hoppe/Köppe-Karkutsch beck.digitax 2024, 46.

22 Vgl. zum Training und Fine-tuning eines LLMs mit rechtlichen Inhalten Colomba et al, SaulLM-7B: A pioneering Large Language Model for Law https://arxiv.org/abs/2403.03883; ferner Glauner LTZ 2024, 24 (26).

23 Siehe zur Retrieval Augmented Generation, RAG, oben unter V...

24 Vgl. zum Begriff "Hybride KI" zB das Informationspapier der Acatech vom 30.11.2023, https://www.acatech.de/publikation/hybride-ki-wissen-und-daten-kombiniert-nutzen/.

25 Vgl. Wagner MMR 2024, 298.

26 Vgl. Paal ZfDR 2024, 129.

Überspitzt gesagt: Dreht sich die Forderung „Ein Mensch muss die Ergebnisse der KI kontrollieren" irgendwann um zur Forderung „Eine KI muss die Ergebnisse des Menschen kontrollieren"?[27]

4. Was wird aus Juristinnen und Juristen?

Wie bedeutend die Veränderungen durch LLMs im Rechtsmarkt letztlich sein werden, lässt sich nicht seriös vorhersagen. Platte Schlagwörter wie *robo-lawyer* oder *robo-judge*[28] sind nicht hilfreich. Im Folgenden der Versuch einer Differenzierung:

a) LLMs erweitern den Zugang zum Recht...

KI-generierte Rechtsauskünfte und Schriftsätze sind für Laien leicht und kostengünstig oder gar kostenlos verfügbar und somit Realität. Solche Auskünfte, auch wenn sie fehlerhaft sein mögen, verbessern für breite Bevölkerungsteile den Zugang zum Recht. In manchen Fallkonstellationen wird dadurch „Rechtsberatung" in Fällen möglich, wo vorher *gar keine* Beratung stattfand, weil sie nicht verfügbar oder zu teuer war. Hier wird somit keine menschliche Beratung ersetzt, sondern auf breiter Basis das rechtliche Niveau „von 0 auf ChatGPT" angehoben.

b) LLMs ersetzen...

In anderen Konstellationen könnte es zu einem Ersatz menschlicher Beratung kommen, wenn eine sofort verfügbare und kostengünstige KI-Auskunft trotz aller Defizite „gut genug" ist. Das kann der Fall sein, wenn die KI-Lösung entweder bereits den Beratungsbedarf befriedigt und Rechtsfrieden herstellt (etwa, wenn die von Laien befragte KI einen Lösungsvorschlag für einen Streit offeriert, den beide Seiten akzeptieren) oder die KI-Lösung aus Sicht der Rechtssuchenden trotz aller Defizite vorzugswürdig ist gegenüber einer Beratung durch Expertinnen oder einem staatlichen Verfahren. In welchem Ausmaß dies künftig der Fall sein wird und welche Implikationen dies nach sich zieht, lässt sich kaum vorhersagen.

c) LLMs unterstützen...

Dort wo eine hohe Qualität der rechtlichen Prüfung erforderlich ist, kann KI nach jetzigem Stand nur unterstützen. Dies lohnt sich dort, wo KI Kosten-, Effizienz- oder Qualitätsvorteile verspricht – etwa bei der Dokumentenanalyse oder Recherche.

Für das juristische Skillset bedeutet dies: In der „Zusammenarbeit" mit KI müssen Juristinnen und Juristen sogar mehr können als in früheren Zeiten, um KI vorteilhaft zu nutzen und so effizienter und/oder qualitativ besser zu arbeiten – der Weg zum *augmented lawyer.* Denn *anders als früher* brauchen sie ein grundsätzliches Verständnis davon, wie LLMs funktionieren, und sie müssen wissen, wie man KI-Systeme richtig bedient und „bepromptet".

Sie müssen zudem in der Lage sein, die Ergebnisse der KI kritisch zu hinterfragen und, wo nötig, zu korrigieren. Sie brauchen also *genau wie früher* die klassischen juristischen Fähigkeiten, um das immer komplexer werdende Recht verstehen und anwenden können. Analytisches Denken, Grundwissen, Methodik und auch Spezialisierung bleiben unentbehrlich.

Die Ausbildung (generell, aber insbesondere auch die juristische – von Universität über Referendariat bis zum Berufseinstieg) steht hier vor großen Herausforderungen. Die wichtigste dürfte sein, die nachfolgende Generation und insbesondere den juristischen Nachwuchs – trotz, aber auch mithilfe von KI – zu eigenständigem, kritischem Denken zu befähigen.[29]

d) ...im Rahmen des Rechts

Es versteht sich von selbst, dass beim KI-Einsatz das geltende Recht beachtet werden muss. Verfassungsgrundsätze, Grundrechte und Verfahrensordnungen setzen dem Einsatz von KI, insbesondere durch Staat und Justiz, Grenzen. Auch private Akteure müssen sich an den Rechtsrahmen halten, den insbesondere Datenschutzrecht, Urheberrecht, Berufsrecht und künftig die KI-Verordnung der EU bilden.

e) Fazit

Insgesamt zeigen die Studien bei aller Vorläufigkeit und offenen methodischen Fragen, dass LLMs die rechtliche Arbeit und den Rechtsmarkt in den kommenden Jahren erheblich verändern werden. Wichtig für Juristinnen und Juristen ist Offenheit. Dazu gehört der Blick über den juristischen Tellerrand sowie eine intensive Auseinandersetzung mit dem Potential, mit den Risiken sowie mit den rechtlichen Rahmenbedingungen. Wir leben in spannenden Zeiten.

---

27 In diese Richtung Risse/Gremminger AnwBl 2022, 24 (26).

28 Vgl. die Fragestellung einer Sachverständigenanhörung des Rechtsausschusses des Landtags Nordrhein-Westfalen vom 13.6.2023: "Besteht die Gefahr, dass Urteile von Richtern und Beschlüsse von Rechtspflegern in Zukunft vollständig durch ChatGPT gefertigt werden und nähern wir uns damit der Gefahr eines „Robo-Jugdes"?

29 Vgl. dazu das Editorial in dieser Ausgabe sowie Heckmann/Rachut, Ordnung der Wissenschaft Heft 2/2024, https://ordnungderwissenschaft.de/wp-content/uploads/2024/03/Heckmann-Druckfahne.pdf.