„KI, LÖS‘ MIR DEN FALL!“ ZUR EVALUIERUNG GROSSER SPRACHMODELLE FÜR ANWENDUNGEN IM DEUTSCHSPRACHIGEN RECHTSWESEN

Bettina Mielke/Christian Wolff

Präsidentin des Landgerichts Ingolstadt, Honorarprofessorin für Rechtsinformatik und Recht der Digitalisierung an der Universität Regensburg, Landgericht Ingolstadt, Auf der Schanz 37, 85049 Ingolstadt, DE, bettina.mielke@lg-in.bayern.de

Lehrstuhl für Medieninformatik, Fakultät für Informatik und Data Science, Universität Regensburg, 93040 Regensburg, DE, christian.wolff@ur.de, https://go.ur.de/mi

Schlagworte: ***Künstliche Intelligenz, generative KI, große Sprachmodelle, Evaluierung, Recht***

Abstract: *Wir diskutieren den möglichen Einsatz großer Sprachmodelle und generativer KI im Rechtswesen und erörtern, wie sich dieser evaluieren lässt. Dazu gehen wir auf bisherige Ansätze zur Evaluierung von großen Sprachmodellen, insbesondere am Beispiel des HELM-Projektes (holistic evaluation of large language models) ein und stellen erste explorative Versuche aus dem deutschsprachigen Raum, die Leistungsfähigkeit großer Sprachmodelle zu bewerten, vor. Auf dieser Basis entwickeln wir Anforderungen an eine systematische Methodik für künftige Evaluierungsstudien.*

1. Einführung

Neben den bekannten und viel diskutierten großen Sprachmodellen wie der GPT-Modellfamilie existiert eine eindrucksvolle Vielfalt kleiner und großer Sprachmodelle, einen guten Überblick dazu geben Minaee et al. 2024. Die Plattform *HuggingFace* (Jones et al. 2024) ist dabei in den letzten Jahren zu einer unverzichtbaren Sammlung von Programmen, Materialien und Daten rund um große Sprachmodelle geworden. Die Anzahl dort gehosteter Ressourcen ist mittlerweile nahezu unübersehbar. Sie lassen sich nach unterschiedlichen Kriterien filtern, unter anderem auch nach Modellen für den juristischen Bereich. Allein zum Stichwort „legal“ finden sich (Stand Anfang Januar 2025) 2.621 *models* und 457 *datasets.* Darunter sind derzeit 18 Modelle, die die Merkmale „german“ und „legal“ erfüllen, alle weisen allerdings vergleichsweise geringe Downloadzahlen auf.[1]

Das deutlich gewachsene Interesse an der KI-Nutzung im Recht zeigt auch der seit 2021 im Rahmen der internationalen Fachtagung zu empirischen Methoden in der automatischen Sprachverarbeitung (*ACL Conference on Empirical Methods in Natural Language Processing*, EMNLP) etablierte Workshop zu *Natural Legal Language Processing*, bei dem bisher 117 Aufsätze erschienen sind, die die ganze Bandbreite der Nutzung großer Sprachmodelle für juristische Anwendungen illustrieren.

2. Große Sprachmodelle im Rechtswesen – internationale Perspektive

Soweit erkennbar, sind bislang nur wenige spezifisch für den rechtlichen Bereich trainierte Modelle öffentlich verfügbar. SaulLM (Colombo et al. 2024), das Anfang 2024 veröffentlicht wurde, dürfte eines der ersten sein. Allerdings liegt hier der Fokus auf Rechtsquellen aus dem angelsächsischen Bereich. Gleichzeitig ist klar,

[1] Vgl. https://huggingface.co/models?language=de&sort=trending&search=Legal.

dass künftig fokussierte und mit hochwertigen Dokumenten aus dem Rechtswesen trainierte Modelle neue Anwendungsmöglichkeiten eröffnen, Fachverlage und Informationsdienstleister haben erste Angebote entwickelt. Lai et al. 2024 geben einen kurzen Überblick zur Entwicklung aktueller Konzepte der künstlichen Intelligenz und deren Anwendung im Rechtswesen. Sie zeigen auf, welche juristischen Anwendungen durch Finetuning großer Sprachmodelle in China seit 2023 entstanden sind und ordnen sie unterschiedlichen Aufgabenbereichen zu. In einem weiteren Überblicksaufsatz nehmen Yang et al. 2024 eine Systematisierung der Nutzung großer Sprachmodelle im Rechtswesen vor, wobei sie zwischen den verwendeten Daten (Sprache, Qualität), den Algorithmen (*adapter tuning, soft prompt tuning, LoRA – low rank adaptation*), unterschiedlichen Aufgaben (*frameworks for resolving different legal tasks*) und Annahmen zur weiteren Entwicklung (*legal document research, case prediction analysis, legal Q & A*) und den damit verbundenen Herausforderungen (Datenqualität, Sicherheit und Datenschutz, Genauigkeit) unterscheiden. Ihre Betrachtung bleibt auf den asiatischen (koreanisch, chinesisch) bzw. den englischsprachigen Raum beschränkt. In einem weiteren Überblicksartikel zur *legal judgement prediction* (LJP) systematisieren Cui/Shen/Wen 2023 LJP-Anwendungen mit 43 Datenkorpora und neun Sprachen, wobei sie neben der Unterscheidung zwischen Zivil-, Straf- und Menschenrechtssachen auch das zugrundeliegende Rechtssystem (kontinentales Recht vs. *common law*) als wesentliches Unterscheidungsmerkmal heranziehen. Die Nutzung existierender generischer LLMs wie ChatGPT z. B. als ChatBot für eine Rechtsberatungsfirma diskutieren Homoki/Ződi 2024 im Rahmen einer Fallstudie.

3. Große Sprachmodelle im Rechtswesen – deutschsprachige Perspektive

Die im deutschsprachigen Raum geführte Diskussion nennt als Einsatzmöglichkeiten generativer KI im Rechtswesen die Klassifikation und die Zusammenfassung von Rechtstexten, die verbesserte Extraktion von Informationen sowie den Bereich der Legal Prediction (vgl. etwa die Zusammenstellungen bei Mielke/Wolff 2023, Rdnr. 584ff; 2024a). Eine Kombination dieser typischen Aufgaben haben Engel/Kruse 2024 im Blick, die einen Prototyp eines KI-generierten Kommentars zu Art. 8 GG auf der Basis eines Korpus mit Entscheidungen des Bundesverfassungsgerichts durch Zusammenfassung der wichtigsten Passagen zu entwickeln versuchen. Dabei kam u.a. ein Prompt mit mehreren Seiten Länge zum Einsatz.

In Niedersachsen nutzt das Projekt MAKI (für **Ma**ssenverfahrensassistenz mithilfe von **KI**) verschiedene große Sprachmodelle, etwa das französische System Mistral (Zander 2024), insbesondere um relevante Informationen aus den Akten zu extrahieren. In ähnlicher Weise sollen LLMs bei der Bearbeitung von Asylverfahren am Verwaltungsgericht Hannover Unterstützung leisten (Irskens/Stock 2024). Beim **E**rkenntnis**mi**ttel**l**assistent EMIL wertet ein großes Sprachmodell eine Datenbank mit relevanten Datenquellen zu den Verhältnissen in verschiedenen Ländern aus. TABEA (**Tatbe**standsassistent für **A**sylverfahren) soll zu einer schnelleren Durchdringung der Akten des Bundesamts für Migration und Flüchtlinge führen, indem u.a. Daten wie Staatsangehörigkeit, Asylantrag, Einreisezeitpunkt und -modalität aufbereitet werden und eine Übersicht der Ausreisegründe bzw. der vorgetragenen Verfolgungshandlungen erstellt werden kann, wobei derzeit kleinere LLMs verwendet werden, um Ressourcen zu schonen (Irskens/Stock 2024, S. 316).

Der Anfang Dezember 2024 publizierte KI-Assistent „Frag den Grüneberg“, den der Verlag C.H. Beck auf der Basis des bekannten BGB-Kommentars entwickelt hat, löste ein starkes Echo in den Medien aus.[2] Beurskens

[2] Vgl. https://presse.beck.de/pressemitteilungen-archiv/2024/frag-den-gruneberg-das-neue-chat-book-von-chbeck/, Zugriff 12 / 2024. Die Reaktionen, die in der Regel auf Einzelbeobachtungen beruhen, variieren stark, vgl. etwa Jochen Zenthöfer in FAZ vom 9. Dezember 2024, der zu einem positiven Fazit kommt. Andere Erfahrungen sind eher negativ, z. B. von Michael Selk (LinkedIn-Beitrag vom 15. Dezember 2024), der sich eine (wirksame) Schönheitsreparaturklausel entwerfen lässt, die nach ganz h. M. als unwirksam einzustufen ist, ähnlich Matthias Kraft (LinkedIn-Beitrag vom 21. Dezember 2024). Differenzierend anhand einzelner Beispiele Stärken und Schwächen herausarbeitend Kay E. Winkler, Markt&Recht vom 12. Dezember 2024. Zu weiteren KI-Tools bei der juristischen Recherche siehe (wenn auch ohne diese kritisch zu hinterfragen) Verena Schillmöller, Schneller, intelligenter, präziser: So verändert Künstliche Intelligenz die juristische Recherche, Fünf neue KI-Tools im Überblick, legal-tech.de vom 4. Juni 2024.

2024 unterzieht ihn einer heuristisch-explorativen Analyse und bewertet anhand einzelner Fragen Kriterien wie Halluzinieren, die Korrektheit und den Umfang der Antworten sowie die differenzierte Darstellung unterschiedlicher rechtlicher Meinungen.

4. Typisierung von Arbeitsaufgaben

In ihrem Überblicksaufsatz zur Nutzung großer Sprachmodelle in „kritischen“ Domänen (Finanz-, Gesundheits- und Rechtswesen) entwickeln Chen et al. (2024) eine Taxonomie von typischen Aufgaben, die mit Hilfe großer Sprachmodelle bewältigt werden können (und geben dazu jeweils Beispiele aus der Literatur):

Legal Question Answering (LQA), also das Beantworten rechtlicher Fragen, wird in einer Reihe von Projekten untersucht; dafür werden spezifische Frage-Antwort-Korpora entwickelt.

Legal Judgment Prediction (LJP) versucht, für spezifische Fälle eine Urteilsvorhersage zu leisten, wobei auch Ansätze für das europäische Recht existieren, wie die Arbeiten von Niklaus zeigen (Niklaus et al. 2023; Niklaus/Chalkidis/Stürmer 2021).

Legal Event Detection (LED) umfasst das Erkennen einschlägiger Hinweise auf Urteile, Normen etc. in Texten, wenngleich hier fraglich ist, ob nicht eine allgemeinere Kategorie (*Legal*) *Information Extraction* oder (*Legal*) *Named Entity Extraction* sinnvoller wäre.

Legal Text Classification (LTC) bedeutet die Klassifikation von Rechtstexten, um z. B. Sachvortrag von rechtlicher Argumentation und Urteilsgründen unterscheiden zu können.

Legal Document Summarization (LDS) beschreibt Systeme, die in der Lage sind, Rechtstexte sinnvoll zusammenzufassen. Eine aktuelle Studie, die sowohl domänenspezifisch trainierte als auch generische große Sprachmodelle auf britische und indische höchstrichterliche Entscheidungen anwenden, stellen Deroy/Ghosh/Ghosh 2024 vor.

Weitere Aufgaben werden als *Other Legal NLP Tasks* zusammengefasst. Homoki/Ződi 2024, S. 447ff weisen darauf hin, dass neben juristischen Kernaufgaben bei der Interaktion mit Fachtext weitere „katalytische“ Einsatzfelder für große Sprachmodelle z. B. im Bereich der Ausbildung entstehen könnten.

5. Große Sprachmodelle und ihre Evaluierung

Spektakuläre Erfolge von ChatGPT und den dahinterstehenden großen Sprachmodellen wurden gefeiert und u. a. mit dem Bestehen des Bar Exam in den USA, wo das neueste getestete Modell, GPT-4, im Schnitt bessere Leistungen erreichte als ein durchschnittlicher Prüfling (Katz/Bommarito/Gao/Arredondo 2024, S. 6), oder mit dem Bestehen des Turing-Tests (*imitation game,* Turing 1950) begründet. Bei letzterem geht es darum, ob Menschen bei der Kommunikation merken, ob sie es mit einer Maschine oder einem Menschen zu tun haben. Der Mensch führt dabei mittels Tastatur und Bildschirm eine Konversation mit zwei verschiedenen Gesprächsteilnehmern, einer Maschine und einem Menschen, und muss anschließend entscheiden, ob er mit einem Menschen oder einer Maschine kommuniziert hat. Nach einer Studie aus dem Jahr 2024 schätzten 54 % der knapp 500 Teilnehmer (Jones/Bergen 2024) das hinter ChatGPT stehende Modell als menschlich ein, 67 % erkannte ihre menschlichen Gesprächspartner (zutreffend) als Menschen, was darauf hindeutet, dass die KI als zunehmend leistungsstark erachtet wird. Der Test betrifft die Fähigkeit, menschliche Gespräche nachzuahmen. Damit wird nichts darüber ausgesagt, ob der Einsatz im Rechtswesen tatsächlich hilfreich ist und zu einer Zeitersparnis oder einer Qualitätssteigerung führt. Dies ist aber entscheidend, gerade angesichts einer aktuellen Umfrage, nach der drei von vier Arbeitnehmern finden, dass KI-Tools ihre Produktivität einschränken und ihre Arbeitsbelastung erhöhen, etwa durch die notwendige Überprüfung von KI-generierten

Inhalten, während gleichzeitig 96 % des leitenden Managements von Produktivitätsgewinnen ausgehen.[3] Auch im Hinblick auf das Bestehen von juristischen Examina ist der Aussagegehalt zu hinterfragen. Vielmehr muss der tatsächliche Nutzen solcher Modelle beim Einsatz im Rechtswesen bewertet werden.

5.1. Holistic Evaluation of Language Models (HELM)

Zur automatischen Evaluierung großer Sprachmodelle sind bereits zahlreiche Ansätze veröffentlicht worden, einer der prominentesten dürfte der in Stanford entwickelte Ansatz HELM (holistische Evaluierung großer Sprachmodelle, *holistic evaluation of language models*) sein (Bommasani/Liang/Lee 2023; Liang et al. 2022): Im Kern besteht die Struktur von HELM aus einer Sammlung von Szenarien für den Einsatz großer Sprachmodelle und Metriken zu ihrer Bewertung. Die Szenarien kombinieren je eine Aufgabe (*task*) mit der Domäne bzw. Datengrundlage (*what*), der betrachteten Autorenschaft der Texte, mit denen trainiert wurde (*who*), ihrem Entstehungszeitraum (*when*) und der untersuchten Sprache. Damit ergibt sich ein flexibler Beschreibungsrahmen. Für die wesentlichen Kernszenarien wurden jeweils die sieben Metriken *accuracy*, *calibration*, *robustness*, *fairness*, *bias*, *toxicity* und *efficiency* bewertet (Bommasani/Liang/Lee 2023, S. 141). So konnten insgesamt 4.900 Evaluierungen von 30 Modellen für 16 Szenarien vorgenommen werden. Durch die Standardisierung von Szenarien und Metriken soll die Bewertung der vielfältigen Modelllandschaft vereinheitlicht werden. Aus der Vielzahl der Einzelversuche lassen sich eine Reihe allgemeiner Erkenntnisse ableiten, z. B. eine grundsätzlich starke Korrelation zwischen *accuracy* und *fairness* und *accuracy* und *robustness* der Modelle (Liang et al. 2022, S. 48, Abb. 24). Mit wenigen Ausnahmen, bei denen eine intellektuelle Bewertung zum Einsatz kann, wurden die Evaluierungen automatisiert durchgeführt. Dass andere Sprachen als Englisch kaum Berücksichtigung finden, ist den Autoren bewusst: “Given this stated taxonomy, we make deliberate decisions on what subset we implement and evaluate, which makes explicit what we miss (e.g. coverage of languages beyond English)” (Liang et al. 2022, S. 3).

5.2. Überblicksstudien zur Evaluierung großer Sprachmodelle

Neben dem umfassenden Ansatz von HELM liegt mittlerweile eine Reihe von Review-Artikeln zur Evaluierung großer Sprachmodelle vor:

Laskar et al. 2024 geben einen Überblick bisheriger Studien, klassifizieren unterschiedliche Herangehensweisen an die Evaluation und schlagen ein dreistufiges Analysemodell für die Bewertung von Evaluationen vor (*evaluation setup, response generation, evaluation methodology*).

Einen ähnlichen Ansatz verfolgen Chang et al. 2024, die zwischen dem ‚what‘, dem ‚where‘ und dem ‚how‘ der Evaluierung unterscheiden. In der Kategorie des ‚what‘ werden sowohl funktionale Fragen der automatischen Sprachverarbeitung als auch konkrete Anwendungsfelder wie *Social Science* (wozu auch das Rechtswesen zählt) zusammengefasst. Hinsichtlich des ‚how‘ unterscheiden sie zwischen automatischer und menschlicher Evaluierung.

Guo et al. 2023 stellen eine weitere Metastudie zur Evaluierung großer Sprachmodelle vor und entwickeln dafür eine Taxonomie mit den Dimensionen Wissen und Fähigkeiten, die Bewertung von Störfaktoren wie Bias oder Toxizität (*alignment evaluation*), die Organisation der Evaluierung (unterschiedliche Typen von Messverfahren) sowie den Aspekt spezialisierter großer Sprachmodelle für spezifische Anwendungsfelder (u.a. „*Legislation“*).

[3] Studie des Dienstleisters Upwork Ltd. auf der Basis einer Umfrage unter 2500 Beschäftigten im April / Mai 2024 in Australien, Großbritannien, Kanada und den USA, vgl. https://investors.upwork.com/news-releases/news-release-details/upwork-study-finds-employee-workloads-rising-despite-increased-c.

Kenthapadi/Sameki/Taly 2024 stellen ein Tutorial zur Evaluierung von LLMs vor, in dem sie die folgenden Dimensionen unterscheiden: Wahrheitstreue, Sicherheit und Anpassung, Bias und Fairness, Robustheit und (technische) Sicherheit, Datenschutz, Vergessen und Urheberrecht, Kalibrierung und Konfidenz / Unsicherheit, Transparenz und kausale Interventionen.

5.3. Einzelstudien zur Evaluierung im Rechtswesen

Neben der Berücksichtigung von Anwendungen im Rechtswesen in den oben vorgestellten Überblicksstudien liegen international bereits zahlreiche Einzelstudien zur Evaluierung des Einsatzes großer Sprachmodelle im Rechtswesen vor, nachfolgend sei eine Auswahl kurz vorgestellt:

Guha et al. 2024 stellen die *LegalBench* vor, die 162 unterschiedliche Aufgaben aus sechs Bereichen juristischer Argumentation (*issue spotting, rule-recall, rule-application, rule-conclusion, interpretation und rhetorical-understanding*) umfasst und mit der sie 20 LLM-Anwendungen im Rechtswesen evaluiert haben. Pipitone/Alami 2024 entwickeln diesen Ansatz weiter, um die Technik des *Retrieval Augmented Generation* (RAG) berücksichtigen zu können. Im Ergebnis wird deutlich, dass die größten kommerziellen Modelle wie GPT-4 die besten Ergebnisse erzielen.

Während sich der Ansatz von Guha et al. 2024 am US-amerikanischen Recht orientiert, präsentieren Li et al. 2024 mit *LexEval* ein umfassendes Bewertungssystem für LLMs im Kontext des chinesischen Rechtswesens, das 23 Aufgaben und 14.150 konkrete Fragen umfasst. Sie bauen dabei auf einer Taxonomie der im Recht benötigten kognitiven Fähigkeiten auf (*legal cognitive ability taxonomy: memorization – understanding – logic inference – discrimination – generation – ethics*, Li et al. 2024, S. 3)

Eine weitere Fallstudie zum US-amerikanischen Steuerrecht stellen Nay et al. 2024 vor, in der sie Multiple-Choice-Fragen mit je einer korrekten Antwort als Evaluationsinstrument zur Bewertung des Outputs großer Sprachmodelle (GPT-3, GPT-3.5, GPT-4) einsetzen. In unterschiedlichen Szenarien und mit / ohne Chain-of-Thought-Prompting kommen die Akkuratheitswerte der Antworten über 60 % nicht hinaus und liegen in der Mehrzahl der Settings deutlich niedriger (Nay et al. 2024, S. 6, 8). Eine aktuelle Studie aus Neuseeland („Better Call GPT“, Martin et al. 2024) vergleicht verschiedene LLMs mit unterschiedlichen Rechtsexperten (*junior / senior lawyer, legal process outsourcer*) hinsichtlich der Fähigkeit, Rechtsprobleme bei der Vertragsdurchsicht zu identifizieren bzw. zu entscheiden, wobei die besten Modelle etwas besser abschneiden als *junior lawyers*, bei allerdings erheblich geringeren Kosten (Kostenreduktion > 99%) bzw. erheblich weniger benötigter Zeit. Als Metriken kommen dabei *recall, precision* und der F1-Score als Kombinationsmaß zum Einsatz.

In ihrer aktuellen Studie zur Halluzinationstendenz kommerzieller LLMs für das (amerikanische) Rechtswesen wie *Lexis+ AI* oder *Westlaw AI-Assisted Research* und *Ask Practical Law AI* kommen Magesh et al. 2024 zu dem Ergebnis, dass diese Modelle bei 17 % bis 33 % der Anfragen halluzinieren. Ähnliche Werte ergeben sich dabei für den Grad der Unvollständigkeit der Antworten. In einer weiteren Studie mit großen Standardmodellen wie GPT-4 werden noch deutlich höhere Werte (75 % und mehr) beobachtet (Dahl/Magesh/Suzgun/Ho 2024).

6. Evaluation großer Sprachmodelle im deutschsprachigen Rechtswesen

Im Kontext der Auseinandersetzung mit den Nutzungsmöglichkeiten im deutschsprachigen Rechtswesen verbleiben Aufsätze, die sich mit den Möglichkeiten, Chancen und Risiken der Anwendung großer Sprachmodelle befassen, selbst wenn sie das Thema Bewertung ansprechen, vielfach auf einer recht allgemeinen Ebene und stellen etwa die Frage, ob es sich bei der Nutzung von KI-Systemen um Rechtsfindung oder Rechtsanwendung handelt (Birkholz 2024), ordnen die Nutzung großer Sprachmodelle aus der Perspektive anwaltlicher Tätigkeit oder des anwaltlichen Berufsrechts ein (Salz/Wiedemann 2024), erörtern den möglichen Einsatz generativer

Modelle in der Justiz (Yuan 2023) oder diskutieren allgemeine rechtliche Rahmenbedingungen der Nutzung großer Sprachmodelle sowie ihre Einsatzfähigkeit in Hochschulen und Verwaltung (Brockmann 2023). Rechtliche Rahmenbedingungen des LLM-Einsatzes mit einem arbeitsrechtlichen Schwerpunkt (menschliche Aufsicht, Rolle von KI-Verordnung und DSGVO) stehen bei Kätscher/Pesch 2024 im Mittelpunkt.

Auch technische Aspekte der LLM-Nutzung werden betrachtet: Das Problem der Halluzinationen wird versucht, durch Anpassung der Prompts in den Griff zu bekommen (Ketteler 2024; Monschau 2024). Die Literatur zur Prompt-Erstellung wird dabei aber nicht oder wenigstens nicht explizit reflektiert (eine umfassende Übersicht zur Prompting-Forschung findet sich bei Schulhoff et al. 2024).

In die Grauzone zwischen Einzelstudie und Wissenschaftssatire fallen Arbeiten wie Bachgrund/Nesum/Bernstein/Burchard 2023, in denen versucht wird, aus ChatGPT wissenschaftlichen Text zum „Pro und Contra für Chatbots in Rechtspraxis und Rechtsdogmatik" zu gewinnen, wozu auch fiktive Autoren wie Lonk Nesum (ein Anagramm zu Elon Musk) eingeführt werden (ähnlich Kraft 2023, aber mit Fokus im pharmazeutischen Bereich).

Anhand eines einzelnen Dialogs mit ChatGPT zur Frage, ob man per Fax fristwahrend Rechtsmittel einlegen kann, arbeitet Herberger 2023 knapp das Problemspektrum (sachlich falsche Auskünfte, Halluzinieren von Gesetzen, mangelnde Aktualität) des Einsatzes generativer KI heraus, dabei handelt es sich aber eher um ein prägnantes Schlaglicht als eine systematische Studie.

Lange 2023 setzt sich inhaltlich sehr detailliert mit der Ausgabe von ChatGPT 3.5 zu einer Anfrage auf dem Gebiet der Testamentsvollstreckung auseinander und kommt dabei zu dem Ergebnis, dass die Antworten „zumeist nicht präzise genug und vereinzelt sogar besorgniserregend falsch" seien, zudem sei das „Erfinden einzelner Normen" zu konstatieren (Lange 2023, S. 570).

Herrlein/Gelück 2023 analysieren Antworten von *Bing AI* (Stand Juni 2023) zum Mietrecht auf ihre Richtigkeit und folgern, dass *ChatGPT* in der Bing-AI-Fassung „im mietrechtlichen Praxistest – der Autoren – gescheitert" sei. Besonders auffällig seien „Widersprüche innerhalb einer Antwort und starke Abweichungen auf Nachfragen", sowie eine eher zufällig anmutende Quellenauswahl. Ihre abschließende Einschätzung lautet: „ChatGPT ist für die mietrechtliche, insbesondere die anwaltliche Praxis derzeit unbrauchbar, jede ungeprüfte Übernahme verbietet sich" (Herrlein/Gelück 2023, S. 519, siehe auch Neuhaus 2023 zum Versicherungsrecht, der Schulnoten auf verschiedene Attribute von 1 bis 6 vergibt und letztlich zu einem unbefriedigenden Gesamtergebnis kommt.).

Die Möglichkeit, Rechtstexte mit generativer KI in einfache Sprache umzuwandeln, untersuchen Mielke/Wolff 2024b und bewerten das mit quantitativen Metriken (Wortanzahl, Textmenge), mit linguistischen Kategorien (syntaktische und lexikalische Unterschiede) und mit Hilfe von Metriken zur Textverständlichkeit.

Für den Anwendungsfall der Anonymisierung diskutieren Adrian/Evert/Heinrich/Keuchen 2024 die Anwendbarkeit klassischer Evaluationsmetriken wie *recall* und *precision* (letztlich *accuracy* im Metrik-Modell von HELM, vgl. auch Adrian 2024, S. 205ff), fordern gleichzeitig aber spezifisch fachlich-juristische Bewertungskriterien. Der Hinweis, dass nicht von vornherein klar ist, welcher *recall*-Wert angemessen ist, dürfte für viele Evaluierungsszenarien gelten.

Dietrich 2024 entwickelt ein eigenes Testszenario, um mit Hilfe unterschiedlicher Prompt-Strategien anhand von je 30 Auslegungs- und zehn Subsumtionsaufgaben für die Rechtsgebiete Zivil-, Straf- und Verwaltungsrecht untersuchen zu können, welche Qualität die von ChatGPT erzeugten Antworten aufweisen. Als Bewertungskriterien verwendet der Autor „Präzision" und „Argumentationstiefe", die er intellektuell auf einer Skala von eins bis fünf Punkten bewertet. Auf die einschlägige Literatur zur Evaluation großer Sprachmodelle und zum Prompting wird dabei nicht Bezug genommen. Nachvollziehen lässt sich in dieser Studie nur der relative Vergleich zwischen den vier Prompting-Strategien, da weder die Methode der Auswahl der Fragestellungen diskutiert wird noch die Metriken definiert und erläutert werden. Insgesamt kommt der Autor zu einem eher zurückhaltenden Fazit. Es handelt sich ungeachtet der benannten Schwächen um eine der systematischeren Studien im deutschsprachigen Bereich.

Engel/Kruse 2024 nehmen eine Bewertung ihres KI-generierten Kommentars zu Art. 8 GG vor und stellen als „enttäuschendste Erfahrung“ fest, dass das Sprachmodell „die Rechtspraxis zwar recht ordentlich erfassen“ könne, sich aber nicht dazu eigne, „die juristischen Nutzer auf Veränderungen in der Auslegung einer Norm hinzuweisen“ (Engel/Kruse 2024, S. 1006). Dies machten die Verfasser daran fest, dass fiktive Entscheidungen zu ebenfalls fiktiven Änderungen der Grundrechtsdogmatik, die sie eingebaut hatten, zwar zitiert würden, aber ohne deutlich zu machen, zu welchen Rechtsänderungen sie geführt haben (Engel/Kruse 2024, S. 1006). Sie vergleichen ihren Prototyp zudem mit verfügbaren Kommentaren zur selben Rechtsmaterie, indem sie die Zahl der Zitate, die Zahl der zitierten Randnummern und den Aktualisierungszyklus auswerten (Engel/Kruse 2024, S. 1006), also auf einer formalen Ebene, die zur Güte nur wenig aussagen kann, da beispielsweise nicht untersucht wird, ob die Zitate zutreffend sind.

7. Anforderungen an die systematische Evaluierung

Nachfolgend machen wir Vorschläge für eine systematische Evaluierung großer Sprachmodelle, die über die generische Bestimmung allgemeiner quantitativer Qualitätsparameter einerseits und die häufig anzutreffende Einzelfallbetrachtung („ich habe Folgendes in ChatGPT ausprobiert“, „Tipps für das Prompten im juristischen Bereich“ (etwa Braegelmann 2023)) andererseits hinausgehen.

Die voranstehende Übersicht zur Evaluierung großer Sprachmodelle zeigt ein zwiespältiges Bild: Auf der einen Seite finden sich einfache quantitative Metriken, die sich (teil-)automatisiert erheben lassen. Diese lassen zwar allgemeine Aussagen zur Ausgabequalität zu, können aber keine inhaltliche Bewertung einer juristischen Auskunft leisten. Auf der anderen Seite sind, insbesondere im deutschsprachigen Bereich, bisherige Analysen an einzelnen Fragestellungen orientiert und gehen, was angesichts der Neuheit der betrachteten Systeme gut nachvollziehbar ist, zunächst explorativ vor.

Wenn man davon ausgeht, dass sich Prozesse und Interaktionsformen im Bereich der juristischen Fachinformation durch dialogfähige KI-Assistenten weiterentwickeln werden, ist die Frage zu stellen, welche Evaluationsmethoden künftig angewandt werden können, um beispielsweise Vergleichsstudien durchzuführen. Das bisherige Evaluationsparadigma für Informationssysteme bzw. im Information Retrieval hat den Relevanzbegriff (Cooper 1971) in den Mittelpunkt gestellt und auf seiner Basis Metriken wie *recall* und *precision* genutzt (Manning/Raghavan/Schütze 2008, S. 137ff). Dabei wurde davon ausgegangen, dass ein Informationssystem als Antwort auf eine Frage eine Reihe von Dokumenten als Treffermenge liefert, die jeweils als relevant oder nicht relevant eingeordnet werden können.

Durch den Einsatz generativer KI-Verfahren ändert sich dieses Szenario grundsätzlich, unbeschadet der Tatsache, dass Information Retrieval-Systeme im Kontext der *Retrieval Augmented Generation* (RAG, vgl. Gao et al. 2023) auch künftig eine wichtige Zulieferfunktion für KI-Modelle innehaben dürften: Künftig steht nicht mehr die zu bewertende Trefferliste als wesentliche Systemausgabe im Mittelpunkt der Evaluierung, sondern die auf die Anfrage (*prompt*) als Text generierte Systemausgabe. Damit erfolgt ein Paradigmenwechsel weg von der Relevanzbewertung für Dokumentmengen hin zur Bewertung von generiertem Text. Auch dies hat als Bewertung von Frage-Antwort-Systemen in der Forschung zu Informationssystemen eine lange Tradition (Liang et al 2022, S. 16f).

Eine künftige Evaluationsmethodik könnte folgende Komponenten enthalten:

1. Nutzung der bei HELM eingeführten Metriken für wesentliche Textmerkmale als Basis.
2. Erfassung und Systematisierung der im juristischen Kontext auftretenden Informationsbedürfnisse, Aufgabenstellungen und Frageformen. Sinnvoll wäre beispielsweise, im Rahmen empirischer Studien die tatsächlich an KI-basierte Dialogsysteme gestellten Fragen zu analysieren, auszuwerten und daraus eine konkrete Typologie zu entwickeln. Erkenntnisse aus der Forschung zu juristischem Informationsverhalten können aufgegriffen werden.[4]

[4] Eine umfangreiche Studie dazu stammt schon aus dem Jahr 1973: Jungjohann/Seidel/Sörgel/Uhlig 1973.

3. Auf dieser Basis erscheint es sinnvoll, ähnlich wie von Dietrich 2024 beschrieben (siehe oben), für unterschiedliche Rechtsgebiete standardisierte Evaluationskataloge aus einschlägigen Fragen zu entwickeln, die die wesentlichen Fragetypen und unterschiedliche Teilgebiete eines Rechtsgebiets abdecken. Mit derartigen Fragekatalogen ließen sich vergleichende Studien besser realisieren.
4. Auch wenn auf eine intellektuelle Bewertung der Ausgaben nicht verzichtet werden kann, könnten die Ausgaben unterschiedlicher Systeme zum selben Rechtsgebiet anhand formaler Kriterien automatisiert verglichen werden (welche Begriffe kommen vor, welche Normen werden zitiert, auf welche Quellen wird zurückgegriffen).
5. Möglich erscheint dabei auch, große Sprachmodelle nicht nur für die eigentliche Aufgabenbearbeitung einzusetzen, sondern als unterstützendes Instrument für den Forschungsprozess, also für die Evaluation zu verwenden: Die unterschiedlichen Ausgaben der verschiedenen juristischen KI-Assistenten könnten mit gezielten Auswertungsaufgaben an ein weiteres KI-System übergeben werden, das eine Auswertung und einen Vergleich vornimmt.

Mit der Einführung einer EU-weiten Regulierung des Einsatzes von Künstlicher Intelligenz durch die europäische KI-Verordnung, die 2024 verabschiedet wurde und die in den kommenden Jahren schrittweise in Kraft treten wird, ergeben sich neue Herausforderungen für die Evaluierung solcher Systeme. Im Hochrisikobereich der KI-Verordnung, worunter bestimmte Bereiche der Rechtspflege und der Strafverfolgung fallen, ist eine Bewertung und Testung erforderlich. Die in den Erwägungsgründen (u. a. in Nr. 59, 60, 66, 74) genannten Dimensionen wie Robustheit und Genauigkeit dürften auch im Sinne der hier betrachteten inhaltlichen Qualitätsbewertung zu verstehen sein und können Bemühungen um die systematische Evaluierung großer Sprachmodelle im Rechtswesen weiteren Vorschub leisten.

Literatur

Adrian, Axel, XAI – Erklärbare Künstliche Intelligenz in der Rechtswissenschaft, NotBZ, 2024, S. 201–212.

Adrian, Axel/Evert, Stefanie/Heinrich, Philipp/Keuchen, Michael, Auslegung des KI-VO-E zur Evaluation von Verfahren der Künstlichen Intelligenz am Beispiel der automatischen Anonymisierung von Gerichtsentscheidungen, in: Sprachmodelle: Juristische Papageien oder mehr? / Language Models: Legal Parrots or more? Tagungsband des 27. Internationalen Rechtsinformatiks Symposions IRIS 2024, S. 85–94.

Bachgrund, Richard/Nesum, Lonk/Bernstein, Max/Burchard, Christoph, Das Pro und Contra für Chatbots in Rechtspraxis und Rechtsdogmatik: Ein kritischer Beitrag zum Auftrag des Rechts und der (Rechts-) Wissenschaft: Argumentieren Sie noch, oder chatten Sie schon?, Computer und Recht, 2023, S. 132–140.

Beurskens, Michael (2024), ‚Frag den Grüneberg', *Legal Tribune Online,* 13.12.2024.

Birkholz, Marco, Mensch oder KI. Rechtsanwendung oder Rechtsfindung?, KIR, 2024, S. 91–94.

Bommasani, Rishi/Liang, Percy/Lee, Tony, Holistic Evaluation of Language Models, Annals of the New York Academy of Sciences, 2023, 1525, S. 140–146.

Braegelmann, Tom (2023), Prompts für ChatGPT – eine Übersicht für Anwälte und Anwältinnen. In: Effizienter arbeiten mit ChatGPT. Potenziale, Prompts und Praxistipps für Kanzleien, Fachinfo-Broschüre 2023, S. 20–23.

Brockmann, Tim, ChatGPT, die Lehre und die Verwaltung – wie verändert KI unsere Institutionen?, NdsVBl, 2023, S. 287–295.

Chang, Yupeng/Wang, Xu/Wang, Jindong/Wu, Yuan/Yang, Linyi/Zhu, Kaijie/Chen, Hao/Yi, Xiaoyuan/Wang, Cunxiang/Wang, Yidong, A Survey on Evaluation of Large Language Models, ACM Transactions on Intelligent Systems and Technology, 2024, S. 1–45.

Chen, Zhiyu Zoey/Ma, Jing/Zhang, Xinlu/Hao, Nan/Yan, An/Nourbakhsh, Armineh/Yang, Xianjun/McAuley, Julian/Petzold, Linda/Wang, William Yang, A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law, arXiv preprint arXiv:2405.01769, 2024.

Colombo, Pierre/Pires, Telmo Pessoa/Boudiaf, Malik/Culver, Dominic/Melo, Rui/Corro, Caio/Martins, Andre FT/Esposito, Fabrizio/Raposo, Vera Lúcia/Morgado, Sofia, SaulLM-7B: A Pioneering Large Language Model for Law, arXiv preprint arXiv:2403.03883, 2024.

Cooper, William S., A Definition of Relevance for Information Retrieval, Information Storage and Retrieval, 1971, S. 19–37.

Cui, Junyun/Shen, Xiaoyu/Wen, Shaochun, A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges, IEEE Access, 2023, S. 102050–102071.

Dahl, Matthew/Magesh, Varun/Suzgun, Mirac/Ho, Daniel E, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, Journal of Legal Analysis, 2024, S. 64–93.

Deroy, Aniket/Ghosh, Kripabandhu/Ghosh, Saptarshi, Applicability of Large Language Models and Generative Models for Legal Case Judgement Summarization, Artificial Intelligence and Law, 2024, S. 1–44.

Dietrich, Henrik, Auslegen und Subsumieren mit ChatGPT, NJW, 2024, S. 2092–2098.

Engel, Christoph/Kruse, Johannes, Kommentar ohne Autor, Können Sprachmodelle das Kommentieren übernehmen?, JZ, 2024, S. 997–1007.

Gao, Yunfan/Xiong, Yun/Gao, Xinyu/Jia, Kangxiang/Pan, Jinliu/Bi, Yuxi/Dai, Yi/Sun, Jiawei/Wang, Haofen, Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv preprint arXiv:2312.10997, 2023.

Guha, Neel/Nyarko, Julian/Ho, Daniel/Ré, Christopher/Chilton, Adam/Chohlas-Wood, Alex/Peters, Austin/Waldon, Brandon/Rockmore, Daniel/Zambrano, Diego, Legalbench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, Advances in Neural Information Processing Systems, 2024, 36.

Guo, Zishan/Jin, Renren/Liu, Chuang/Huang, Yufei/Shi, Dan/Yu, Linhao/Liu, Yan/Li, Jiaxuan/Xiong, Bojian/Xiong, Deyi, Evaluating Large Language Models: A Comprehensive Survey, arXiv preprint arXiv:2310.19736, 2023.

Herberger, Marie, Gesprächs- und Kooperationspartner der besonderen Art für die Anwaltsarbeit: ChatGPT und die Folgen, ZAP, 2023, S. 465–466.

Herrlein, Jürgen/Gelück, Johannes, ChatGPT im mietrechtlichen Praxistest (aus Anwaltssicht): „Wundermittel“ oder doch „Rohrkrepierer“?, NZM, 2023, S. 513–519.

Homoki, Péter/Ződi, Zsolt, Large Language Models and their Possible Uses in Law, Hungarian Journal of Legal Studies, 2024, S. 435–455.

Irskens, Gesine/Stock, Niclas, KI-Einsatz an den Verwaltungsgerichten – ein Gamechanger?, DRiZ, 2024, S. 314–317.

Jones, Cameron R/Bergen, Benjamin K, People Cannot Distinguish GPT-4 from a Human in a Turing Test, arXiv preprint arXiv:2405.08007, 2024.

Jones, Jason/Jiang, Wenxin/Synovic, Nicholas/Thiruvathukal, George K/Davis, James C, What do we Know about Hugging Face? A Systematic Literature Review and Quantitative Validation of Qualitative Claims, arXiv preprint arXiv:2406.08205, 2024.

Jungjohann, Knut/Seidel, Ulrich/Sörgel, Werner/Uhlig, Sigmar, Informationsverhalten und Informationsbedarf von Juristen, Teil 1: Analyse-Band. Eine Erhebung von Infratest Sozialforschung, München, im Auftrag des Bundesministeriums der Justiz und der Gesellschaft für Mathematik und Datenverarbeitung, 1973 [Reprint 2020].

Kätscher, Johannes/Pesch, Paulina Jo, Automatisierte Entscheidungsfindung mittels großer Sprachmodelle (LLM) im Beschäftigtenkontext. Interview mit einem Chatbot, KIR, 2024, S. 46–55.

Katz, Daniel Martin/Bommarito, Michael James/Gao, Shang/Arredondo, Pablo, GPT-4 Passes the Bar exam, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2024, S. 20230254.

Kenthapadi, Krishnaram/Sameki, Mehrnoosh/Taly, Ankur, Grounding and Evaluation for Large Language models: Practical Challenges and Lessons Learned (Survey). Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, S. 6523–6533.

Ketteler, Marion, Maschinenlernen: KI-Sprachmodelle im Kanzleieinsatz, KP, 2024, S. 65–71.

Kraft, Matthias, Die Erfindung von Emozipnal, Kraft Verlag, ein Imprint der Efficient Publishing UG, Mühldorf am Inn 2023.

Lai, Jinqi/Gan, Wensheng/Wu, Jiayang/Qi, Zhenlian/Yu, Philip S., Large Language Models in Law: A Survey, AI Open, 2024.

Lange, Knut Werner, Pflichten des Erben bei angeordneter Testamentsvollstreckung – Was meint ChatGPT dazu?, ZEV, 2023, S. 565–570.

Laskar, Md Tahmid Rahman/Alqahtani, Sawsan/Bari, M Saiful/Rahman, Mizanur/Khan, Mohammad Abdullah Matin/Khan, Haidar/Jahan, Israt/Bhuiyan, Amran/Tan, Chee Wei/Parvez, Md Rizwan, A Systematic Survey and

Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations, arXiv preprint arXiv:2407.04069, 2024.

Li, Haitao/Chen, You/Ai, Qingyao/Wu, Yueyue/Zhang, Ruizhe/Liu, Yiqun, LexEval: A Comprehensive Chinese Legal Benchmark for Evaluating Large Language Models, arXiv preprint arXiv:2409.20288, 2024.

Liang, Percy/Bommasani, Rishi/Lee, Tony/Tsipras, Dimitris/Soylu, Dilara/Yasunaga, Michihiro/Zhang, Yian/Narayanan, Deepak/Wu, Yuhuai/Kumar, Ananya, Holistic Evaluation of Language Models, arXiv preprint arXiv:2211.09110, 2022.

Magesh, Varun/Surani, Faiz/Dahl, Matthew/Suzgun, Mirac/Manning, Christopher D/Ho, Daniel E, Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, arXiv preprint arXiv:2405.20362, 2024.

Manning, Christopher D/Raghavan, Prabhakar/Schütze, Hinrich, Introduction to Information Retrieval, Cambridge University Press, Cambridge 2008.

Martin, Lauren/Whitehouse, Nick/Yiu, Stephanie/Catterson, Lizzie/Perera, Rivindu, Better Call GPT, Comparing Large Language Models against Lawyers, arXiv preprint arXiv:2401.16212, 2024.

Mielke, Bettina/Wolff, Christian, Künstliche Intelligenz und Large Language Models in der Rechtsprechung, LRZ – E-Zeitschrift für Wirtschaftsrecht und Digitalisierung, 2023, https://www.lrz.legal/2023Rn560.

Mielke, Bettina/Wolff, Christian, Maschinelle Lernverfahren als KI-Komponenten in Digitalisierungsprojekten der Justiz, LTZ, 2024a, S. 144–153.

Mielke, Bettina/Wolff, Christian, Verständliche Rechtstexte mit Hilfe grosser Sprachmodelle?, in: Sprachmodelle: Juristische Papageien oder mehr? / Language Models: Legal Parrots or more? Tagungsband des 27. Internationalen Rechtsinformatiks Symposions IRIS 2024, 2024b, S. 27–37.

Minaee, Shervin/Mikolov, Tomas/Nikzad, Narjes/Chenaghlu, Meysam/Socher, Richard/Amatriain, Xavier/Gao, Jianfeng, Large Language Models: A Survey, arXiv preprint arXiv:2402.06196, 2024.

Monschau, Norbert, Künstliche Intelligenz: Einsatzmöglichkeiten von ChatGPT & Copilot in der erbrechtlichen Praxis EE, 2024, S. 27–31.

Nay, John J/Karamardian, David/Lawsky, Sarah B/Tao, Wenting/Bhat, Meghana/Jain, Raghav/Lee, Aaron Travis/Choi, Jonathan H/Kasai, Jungo, Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence, Philosophical Transactions of the Royal Society A, 2024, S. 20230159.

Neuhaus, Kai-Jochen, Künstliche Intelligenz im Versicheungsrecht – grundsätzliche Überlegungen und praktische Erfahrungen am Beispiel von ChatGPT und der Berufsunfähigkeitsversicherung, VersR, 2023, 1401–1415.

Niklaus, Joel/Chalkidis, Ilias/Stürmer, Matthias, Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark, arXiv preprint arXiv:2110.00806, 2021.

Niklaus, Joel/Matoshi, Veton/Stürmer, Matthias/Chalkidis, Ilias/Ho, Daniel E, Multilegalpile: A 689gb Multilingual Legal Corpus, arXiv preprint arXiv:2306.02069, 2023.

Pipitone, Nicholas/Alami, Ghita Houir, LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain, arXiv preprint arXiv:2408.10343, 2024.

Salz, Konstantin /Wiedemann, Max-Julian, KI im Mandat – Chancen und rechtliche Grenzen, NJW, 2024, S. 1634–1636.

Schulhoff, Sander/Ilie, Michael/Balepur, Nishant/Kahadze, Konstantine/Liu, Amanda/Si, Chenglei/Li, Yinheng/Gupta, Aayush/Han, HyoJung/Schulhoff, Sevien, The Prompt Report: A Systematic Survey of Prompting Techniques, arXiv preprint arXiv:2406.06608, 2024.

Turing, Alan M., Computing Machinery and Intelligence, Mind, 1950, S. 433–460.

Yang, Xiaoxian/Wang, Zhifeng/Wang, Qi/Wei, Ke/Zhang, Kaiqi/Shi, Jiangang, Large Language Models for Automated Q&A Involving Legal Documents: A Survey on Algorithms, Frameworks and Applications, International Journal of Web Information Systems, 2024, S. 413–435.

Yuan, Tianyu, Justiz GPT: Möglichkeiten und Grenzen des Einsatzes generativer Sprachmodelle bei gerichtlichen Entscheidungen, LTZ, 2023, S. 195–202.

Zander, Henning, Mit KI gegen den Aktenstau, AnwBl vom 30. April 2024, https://anwaltsblatt.anwaltverein.de/de/themen/markt-chancen/ki-projekte-justiz (letzter Zugriff 1 / 2025).