THOMAS HOEREN/MERLIN ROMBACH

Von Token zu Text: Möglichkeiten und Grenzen individualisierbarer KI-Assistenten in der Rechtspraxis

„Fasten your seatbelts!" – mit dieser charakteristischen Sentenz pflegte Gerald Spindler seine Vorträge zu eröffnen. Diese Aufforderung wirkt geradezu prophetisch für die rasante Entwicklung der künstlichen Intelligenz (KI), deren rechtliche wie praktische Implikationen ihn in den letzten Jahren seines akademischen Wirkens intensiv beschäftigten. In einem 2019 publizierten Editorial mahnte Spindler zur kritischen Reflexion der vermeintlichen „Intelligenz" von KI-Systemen, plädiert aber auch für pragmatische Lösungsansätze zur Bewältigung der mit der Implementierung eingehergehenden juristischen Herausforderungen.[1] Statt eines vorschnellen legislativen Eingreifens und Überregulierung sprach er sich in einem Aufsatz schon vor fast zehn Jahren dafür aus, die rasanten Entwicklungen künstlicher Intelligenz zuvörderst mit den bewährten Mitteln des Rechts anzugehen.[2]

Die juristische Praxis sieht sich heute mit einer Vielzahl von KI-gestützten Systemen konfrontiert, deren Einsatz zwar erhebliche Potenziale verspricht, zugleich aber grundlegende praktische und rechtliche Fragen aufwirft. In Fortentwicklung Spindlers bemerkenswerter Fähigkeit, juristisch komplexe Sachverhalte mit profundem technischem Sachverstand zu verbinden, soll im Folgenden anhand eines konkreten Forschungsprojektes exemplarisch aufgezeigt werden, wie der niederschwellige Einsatz von KI-Systemen in der Rechtspraxis gestaltet werden kann. Ausgehend von einer technischen Einführung in KI- und insbesondere Chatbot-Systeme wird der ITM-GPT als Fallstudie vorgestellt, der als Arbeitshilfe im deutschen IT-Recht konzipiert wurde. Schon die Entwicklung und Nutzung herkömmlicher KI-Systeme geht mit unzähligen – oftmals ungeklärter – Rechtfragen einher. Eine vertiefte Auseinandersetzung mit den hinzukommenden Rechtsfragen individualisierbarer KI-Systeme würden den Rahmen dieser Ausarbeitung – die zuvörderst technische wie praktische Aspekte in den Fokus nimmt – übersteigen. Die folgenden Ausführungen verstehen sich daher vielmehr als initialer Impuls für eine breiter angelegte rechtliche Diskussion, welche die oftmals leider wenig Beachtung findenden technischen Grundlagen und praktischen Einsatzszenarien nicht unberücksichtigt lassen sollte.

I. KI-Systeme zwischen technischer Realität und juristischer Konzeption

Bereits bei einer ersten Annäherung an KI-Systeme stellt sich ein grundlegendes Problem: Die Vielfalt und Komplexität dieser Systeme erschwert eine einheitliche definitorische Bestimmung und resultiert in einer Vielzahl möglicher Anknüpfungspunkte, welche sowohl rechtliche, technische sowie funktionale Aspekte umfassen.[3] Exemplarisch werden nachfolgend zwei Definitionsansätze präsentiert. Sodann erfolgt eine taxonomische Einordnung von Chatbots als derzeit prominenteste Erscheinungsform von KI-Systemen.

---

[1] Spindler IIC 2019, 1049 (1049, 1051).

[2] Spindler CR 2015, 766 (774).

[3] Ebers, StichwortKommentar Legal Tech/Yuan, 2023, KI Rn. 11; Herberger NJW 2018, 2825 (2826).

1. Definitionsansätze künstlicher Intelligenz

Während sich bisher schon bei der grundlegenden Begriffsbestimmung eine bemerkenswerte Divergenz der verschiedenen Definitionsansätze zeigte, gibt Art. 3 Nr. 1 Verordnung (EU) 2024/1689 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz[4] (iF KI-VO) erstmals eine feststehende europarechtliche Definition vor. Danach ist ein KI-System

„ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können"

Bei näherer Betrachtung verdeutlicht diese Definition deutlich den legislativen Spagat zwischen dem Streben nach Technologieneutralität einerseits und dem Erfordernis hinreichender rechtlicher Bestimmbarkeit andererseits. In enger Anlehnung an die überarbeitete OECD-Definition[5] hat sich der europäische Gesetzgeber dabei für einen sehr weiten Definitionsansatz entschieden, der jedoch nicht zuletzt deshalb in seiner praktischen Abgrenzungsfunktion erhebliche Defizite aufweist.[6] Die fehlende Unterscheidungskraft zwischen KI- und konventionellen Software-Systemen zeigt sich bereits in der Grundvoraussetzung des „maschinengestützten Systems", die letztlich jegliche Form computerbasierter Datenverarbeitungen umfasst.[7] Ähnliches ergibt sich auch für die graduell abstufbare Autonomie („in unterschiedlichem Grade autonomen Betrieb"), den von der Definition beschriebenen Outputs und dem ubiquitären Merkmal physischer oder virtueller Beeinflussung der Umgebung, welche allesamt durch simple regelbasierte Programmierung erreicht werden können. In der Folge wird nahezu jedes Softwaresystem die Definitionsmerkmale erfüllen können.[8]

Doch auch über diese vergleichsweise junge normative Annäherung hinaus zeigt sich bei der definitorischen Grundlegung eine Heterogenität der wissenschaftlichen Ansätze. Zusammenfassen lassen sich die Ansätze durch die vier zentralen Dimensionen des menschlichen, rationalen, Handelns und Denkens. Die Begriffe des Denkens und Handelns sind dabei in einem übertragenen technischen Sinne zu verstehen und implizieren gerade nicht das Vorliegen eines eigenen freien Willens oder die Fähigkeit zur autonomen Entscheidung im Sinne juristischer Handlungslehre.[9] Vielmehr geht es um eine Approximation menschlicher Verhaltensweisen. Exemplarisch stellt der berühmte, auf Alan Turing zurückgehende Turing-Test primär auf die Dimension des menschlichen Handelns ab.[10] Der Test beruht auf der Idee, dass eine Maschine dann als „intelligent" einzustufen ist, wenn sie im Rahmen

---

[4] Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13.6.2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz), ABl. L, 2024/1689.

[5] OECD (2024), Explanatory memorandum on the updated OECD definition of an AI system, OECD Artificial Intelligence Papers, No. 8, OECD Publishing, S. 4 f.

[6] Wendehorst/Nessler/Aufreiter/Aichinger MMR 2024, 605 (605); Bomhard/Siglmüller RDi 2024, 45 (45).

[7] Eine Abgrenzung wird damit allenfalls gegenüber biologischen Systemen erreicht, dazu: Wendehorst/Nessler/Aufreiter/Aichinger MMR 2024, 605 (607).

[8] Krit. zur Begriffsbestimmung nach der KI-VO: Wendt/Wendt, Das neue KI-Recht, 1. Aufl. 2024, § 2 Rn. 10; Hacker/Berz ZRP 2023, 226 (227); Roos/Weitz MMR 2021, 844 (845); Wendehorst/Nessler/Aufreiter/Aichinger MMR 2024, 605 (606 ff.).

[9] Ebers, StichwortKommentar Legal Tech/Yuan, 2023, KI Rn. 11.

[10] A. M. Turing, Computing Machinery and Intelligence, Mind LIX (1950), 236, 433–460; Turing selbst nannte den Test noch das „Imitation Game". Über die Zeit wurde der Test simplifiziert und unter dem heutigen Namen „Turing-Test" bekannt. Ausf. dazu: Sesink, Menschliche und künstliche Intelligenz. Der kleine Unterschied, 2012, Kap. 3, S. 27 ff.

einer Konversation einen Menschen dazu verleiten kann, zu glauben, dass sie selbst ein Mensch sei.[11] Im Gegensatz zu anderen Begriffseingrenzungen ist der Turing-Test dabei aber ein rein funktionaler Ansatz, der nicht auf die innere Funktionsweise oder konkrete technische Implementation, sondern allein auf das nach außen sichtbare Verhalten abstellt. Durch die dem Test innewohnende Beschränkung auf kommunikationsbasierte Anwendungen eignet er sich damit wenig als Klassifizierungsinstrument nicht interaktiver, lediglich im Hintergrund operierender Systeme.[12]

2. Taxonomie moderner Chatbot-Systeme

Für die im Alltag präsenteste Form und im Fokus dieses Beitrages stehende künstlicher Intelligenz – generative Chatbots wie ChatGPT, Claude.ai, LLaMA, Bard, PalM und Co. – gibt der vorbeschriebene Turing-Test hingegen taugliche Anforderungen vor, die eine Differenzierung gegenüber herkömmlichen, nicht KI-basierten Chatbots ermöglichen.

Chatbots sind Softwareapplikationen zur Simulation menschlicher Konversationen mittels natürlicher Sprache. Ihre taxonomische Einordung auf Grundlage des Turing Tests kann dabei einer häufigen Misskonzeption entgegenwirken, wonach Chatbots oft vorschnell mit künstlicher Intelligenz in Verbindung gebracht werden. Auf vordefinierte Wenn-Dann-Regeln (regelbasierte Chatbots) oder dem Abruf vorgefertigter Antworten basierende Systeme (retrieval-basierte Chatbots)[13] sind nicht in der Lage, die für menschliche Konversation charakteristische sprachliche Flexibilität und Kontextanpassung zu simulieren (Natural Language Processing).[14] Dies zeigt sich gerade dann, wenn die Konversation den engen Rahmen der programmierten Regeln und Stichworte oder des hinterlegten Antwortkataloges verlässt. Hier fehlt es den Systemen sowohl an der Fähigkeit zur eigenständigen Aufnahme und Strukturierung von Wissen (Knowledge Representation) als auch an der Kompetenz, aus diesen Informationen logische Schlussfolgerungen zu ziehen (Automated Reasoning).[15] Jedenfalls die fehlende Möglichkeit aus Erfahrungen zu lernen und das eigene Verhalten entsprechend anzupassen wird dabei zumeist schon nach wenigen Interaktionszyklen zu einer Durchbrechung der für den Turing-Test notwendigen menschlichen Approximation führen und die maschinelle Natur offenbaren.

Die inzwischen weit verbreiteten generativen KI-Chatbots kombinieren dagegen die vorbeschriebenen Merkmale und Fähigkeiten zu leistungsfähigen Systemen, die menschenähnlich in Wort, Bild und Sprache mit Menschen interagieren können. Den Turing Test vollständig bestehen können sie gleichwohl nicht unbedingt.[16]

II. Von Token zu Text – Funktionsweise von LLMs

Die im späteren Verlauf erläuterten Limitierungen und Anwendungsrisiken generativer Chat-bots ergeben sich zuvörderst aus deren technischer Funktionsweise, weshalb ein zumindest rudimentäres technisches Verständnis der zugrundeliegenden Large Language Models (LLMs) bei der Nutzung vorteilhaft ist.

---

[11] A. M. Turing, Computing Machinery and Intelligence, Mind LIX (1950), 236, 434.

[12] Dahingehend: Sesink, Menschliche und künstliche Intelligenz. Der kleine Unterschied, 2012, Kap. 3, S. 37 f.

[13] Chatbots dieser Art sind so programmiert, dass sie zu Beginn auf vorgegebene, gezielte Anfragen antworten. Für die Benutzer besteht dabei nur eine begrenzte Anzahl an Eingabemöglichkeiten: Reda Global Scientific Journals 12 (2024) 10, 1617 (1619).

[14] Reda Global Scientific Journals 12 (2024) 10, 1617 (1620).

[15] Zu den einzelnen Merkmalen vgl.: Ebers, StichwortKommentar Legal Tech, 2023, KI Rn. 11.

[16] Wendt/Wendt, Das neue KI-Recht, 1. Aufl. 2024, § 2 Rn. 1.

LLMs sind mit umfangreichen Textdatensätzen trainierte künstliche Intelligenzmodelle, die menschliche Sprache verstehen und generieren können.[17] Viele der heute gängigen Modelle sind dabei sog. Multimodal Large Language Models (MLLMs), die neben Text auch andere Eingabeformen wie Bild und Ton verarbeiten können.[18] Der KI-Boom der letzten Jahre – bei dem Chat-KIs im Vordergrund der öffentlichen Wahrnehmung stehen – wurde primär durch erhebliche Fortschritte sog. Multi-Head-Attention-Mechanismen, insbesondere der Entwicklung der Transformer-Architektur, erzielt.

1. Transformer als technisches Fundament

Die Nutzung von Transformern unterscheidet sich wesentlich von früheren Ansätzen der Sprachverarbeitung durch die Implementierung eines Selbstaufmerksamkeitsmechanismus (engl.: Self-Attention).[19] Damit das Modell natürliche Sprache verarbeiten kann, erfolgt zunächst eine Tokenisierung, also eine Zerlegung in maschinenlesbare Einheiten (sog. Token). Texte werden dabei feingranular in einzelne Wörter, Wortteile oder andere semantische Einheiten zerlegt, um dem Modell eine präzise Erfassung des Inhaltes und der Zusammenhänge und sprachlichen Nuancen zu ermöglichen.[20]

Während ältere Sprachmodell die Texte im Anschluss an die Tokenisierung architekturbedingt lediglich sequenziell verarbeiten konnten, ermöglicht der Selbstaufmerksamkeitsmechanismus eine simultane Berücksichtigung aller anderen Token einer Sequenz. Neben einer effizienteren Verarbeitung können dabei insbesondere Zusammenhänge über längere Textpassagen hinweg besser verstanden werden.[21] Durch den Einsatz von Multi-Head-Attention kann das Modell dabei zusätzlich mehrere Self-Attention-Operationen parallel durchführen und die Informationsverarbeitung weiter optimieren und gleichzeitige verschiedene Perspektiven analysieren.[22]

2. Probabilistische Sprachmodellierung

Die fortschrittlichen Chat Bots zeichnen sich neben der Fähigkeit des vermeintlichen Textverständnisses insbesondere durch ihre Fähigkeit aus, grammatikalisch und zumeist sachlich zutreffende, vom jeweiligen Kontext und der Nutzeranfrage abhängige Texte zu generieren. Allerdings haben moderne LLMs kein Sprachverständnis im menschlichen Sinne, vielmehr basiert diese Fähigkeit auf der sog. Next Token Prediction.[23] Das Modell berechnet auf Grundlage seiner mehrschichtigen Transformer-Architektur für jeden nächsten Token eine bedingte Wahrscheinlichkeit auf Basis des bisherigen Kontextes. Grundlage der Wahrscheinlichkeitsberechnung sind Muster, die das Modell aus den Trainingsdaten gelernt hat und die als Parameter im Modell inkorporiert sind.[24] Soll ein Modell beispielsweise den Satz „Die Hauptstadt von Frankreich ist …" vervollständigen, werden die Wahrscheinlichkeiten

---

[17] Deligiannidis/Dimitoglou/Arabnia, Artificial Intelligence/Choi/Jo, 2024, Vol. 1, S. 281.

[18] Glauner LTZ 2024, 24 (32).

[19] Wegweisend war in diesem Zusammenhang das von Google-Research veröffentlichte Paper von Vaswani et. al., Attention Is All You Need, 31st Conference on Neural Information Processing Systems (2017).

[20] Turner, An Introduction to Transformers, Feb. 2024, https://doi.org/10.48550/arXiv.2304.10557, S. 1.

[21] Glauner LTZ 2024, 24 (30 f.).

[22] Turner, An Introduction to Transformers, Feb. 2024, S. 3.

[23] Zu diesem Begriff: Bachmann/Nagarajan, The Pitfalls of Next-Token Prediction, https://doi.org/10.48550/arXiv.2403.06963, S. 3.

[24] Turner, An Introduction to Transformers, Feb. 2024, S. 3.; beim inzwischen veralteten GPT 3.0, sog. Text-Da-Vinci-003 Modell etwa 175 Mrd. Parameter: Singh et. al., CODEFUSION: A Pre-trained Diffusion Model for Code Generation, 2023 Conference on Empirical Methods in Natural Language Processing, S. 11700.

für mögliche Folgetoken berechnet, wobei „Paris" im Vergleich zu anderen möglichen Antworten (Berlin, London usw.) die höchste Wahrscheinlichkeit aufweisen und vom Modell als zutreffend ausgewählt werden wird.

Die Qualität und gewisse Charakteristika der generierten Texte können dabei durch verschiedene Parameter gesteuert werden. Anpassungen haben dabei insbesondere Einfluss auf den Auswahlprozess der Token. So ermöglicht etwa die Anpassung der Temperatur eine gezielte Adjustierung zwischen präziseren, aber möglicherweise monotoneren Antworten auf der einen, oder kreativeren, aber potenziell weniger akkuraten Ausgaben auf der anderen Seite.[25] Diese probabilistische Herangehensweise und Anpassungsmöglichkeiten durch Parameter und Kontext führen im Ergebnis dazu, dass moderne Chat Bots nicht deterministische, im klassischen Sinne berechnete Ergebnisse produzieren. Eine zentrale Charakteristik dieser Systeme ist daher auch, dass die von unzähligen Parametern abhängigen statistischen Wahrscheinlichkeitsberechnungen zu einer hohen Antwortvarietät führen – selbst bei sehr ähnlichen oder identischen Eingaben.

3. KI-Training

Grundlage des vorbeschriebenen Sequence Modeling bildet ein komplexer Trainingsprozess. Es erfolgt zunächst ein Pretraining, bei welchem die Transformer und Attention-Heads darauf konditioniert werden, sprachliche Muster und Strukturen in umfangreichen Textkorpora erkennen und reproduzieren zu können. Die sehr umfangreichen Trainingsdaten werden als Eingabetexte ebenfalls tokenisiert und die Analyse dieser führt zu einer erst initialen, anschließend iterativen Anpassung der Modellparameter. Auf Grundlage der Trainingsdaten werden dem System unzählige sog. Vorhersageaufgaben gestellt, bei dem das Modell selbstständig bestimmte Tokensequenzen vorhersagen muss und dabei die Parameter auf Basis der bei den Vorhersagen entstehenden Fehlern optimiert.[26]

Nach Abschluss dieses initialen Pretrainings erfolgt zumeist ein gezieltes Fine-Tuning, etwa durch Instruction-Tuning oder Reinforcement Learning from Human Feedback (RLHF), wodurch die grundlegenden sprachlichen und fachlichen Fähigkeiten des Modells zusätzlich an spezifische Anwendungskontexte und Qualitätsanforderungen angepasst werden.[27] Diese mehrstufige Trainingsmethodik ermöglicht den Systemen letztlich ausreichend präzise Parameter zu entwickeln, sodass für unterschiedliche Szenarien zutreffende Tokenfolgen wahrscheinlichkeitsbasiert generiert werden können.

III. Custom GPTs: Möglichkeiten und Grenzen individualisierbarer KI-Assistenten am Beispiel des ITM-GPT

Die zunehmende Verfügbarkeit generativer künstlicher Intelligenz wirft die Frage auf, inwieweit diese Systeme tatsächlich sinnvoll für die Rechtspraxis nutzbar gemacht werden können. Neben einer Vielzahl für juristische Anwendungsfälle spezialisierter – regelmäßig aber sehr kostspieliger – KI-Systeme bieten einige Anbieter inzwischen vielversprechende Individualisierungsmöglichkeiten, mit denen die Standardmodelle auch von technischen Laien für spezifische Anwendungsfälle angepasst werden können. So ermöglicht die Einführung der Custom GPTs (kurz: GPTs) durch OpenAI das vermeintlich

---

[25] Die Temperatur ist ein sog. Inference Parameters, mit dem das Verhalten eines LLMs wesentlich beeinflusst werden kann: O. Campesato, Large Language Models: An Introduction, 2024, S. 167 f.

[26] Ausf. zum Pre-Training vgl.: Zhao et. al., A Survey of Large Language Models, https://doi.org/10.48550/arXiv.2303.18223, S. 16 ff.; Glauner LTZ 2024, 24 (26 f.).

[27] O. Campesato, Large Language Models: An Introduction, 2024, S. 332.

niederschwellige Anpassen der Grundfunktionalitäten von GPT4 an spezifische Anwendungsfälle.[28]

1. Grundlagen und Funktionsweisen

Zunächst dürfen die Individualisierungsmöglichkeiten mittels Custom GPTs nicht mit den Anpassungsmöglichkeiten von KI-Modellen über sog. Application Programming Interfaces (APIs) verwechselt werden. Über eine solche Schnittstelle kann ein Fine-Tuning des zugrundeliegenden Modells, beispielsweise durch zusätzliche Trainingsdaten, vorgenommen werden. Während eine API-Integration damit eine tiefgreifende technische Modifikation eines Modells ermöglicht, ist sie aber auch mit erheblichem technischem Aufwand verbunden und erfordert eine tiefergehende technische Expertise.[29]

Dagegen beschränkt sich der Custom-GPT-Ansatz auf eine eher oberflächliche Verhaltenssteuerung. Durch die Definition von Verhaltensanweisungen (Instructions) können im ersten Schritt Vorgaben für die Interaktion mit den Nutzern formuliert werden. Sie dienen dazu, den Kommunikationsstil und die inhaltliche Ausrichtung der Interaktion an spezifische Anforderungen anzupassen. So kann die Sprache formal oder informell gehalten, ein thematischer Fokus gesetzt oder unerwünschte Konversationsrichtungen ausgeschlossen werden.[30] Im technischen Sinne sind diese Anweisungen jedoch stets nur temporär und werden – anders als beim Modelltraining oder Feintuning – nicht dauerhaft in die Modellparameter integriert; sie wirken vielmehr als zusätzlicher Kontext. Ihr Einfluss ist dadurch begrenzt. Sie wirken als dem Nutzerprompt vorangestellt Anweisungen lediglich oberflächlich und können das Verhalten des zugrundeliegenden Sprachmodells nicht nachhaltig verändern.

In einem zweiten Schritt ermöglicht die Einbindung einer kontextuellen Wissensbasis die Bereitstellung zusätzlicher Informationen, die nicht in den vorab trainierten Daten erhalten sind. Allerdings erfolgt auch hierbei keine permanente Integration der neuen Informationen in die Modellparameter. Stattdessen werden relevante Daten aus der Wissensbasis in Echtzeit und in Abhängigkeit des aktuellen Verarbeitungskontextes abgerufen und in die Antworten eingebunden. Damit kann der Custom-GPT mit aktuellem oder spezifischem Wissen auf Anfragen reagieren. Da es aber auch hier gerade nicht zu einem Neutraining kommt, sondern die externen Daten lediglich als eine Art Referenzmaterial dienen, bleibt die grundlegende Struktur und das Verhalten des Modells auch durch die Integration eigenen Wissens unverändert.

2. Praktische Bedeutung am Beispiel des ITM-GPT

Um die praktische Bedeutung dieser niederschwellig anpassbaren KI-Modelle zu untersuchen, haben die Autoren im Rahmen eines Forschungsprojektes am Institut für Informations-, Telekommunikations- und Medienrecht (ITM) der Universität Münster den Custom-GPT IT-Verträge – der ITM GPT entwickelt und erprobt.[31]

Der ITM-GPT wurde dabei mit dem recht ambitionierten Zielkatalog konzipiert, Laien sowohl bei der Erstellung von IT-Verträgen zu unterstützen als auch allgemeine Fragen zum IT-Recht beantworten zu können und auch bestehende Vertragswerke einer kritischen Analyse zu unterziehen. Diese Anforderungen sind vor dem Hintergrund der besonderen Kom-

[28] OpenAI, Introducing GPTs, https://openai.com/index/introducing-gpts/ (abgerufen 13.12.2024).

[29] https://platform.openai.com/docs/guides/fine-tuning (abgerufen 13.12.2024).

[30] Mit verschiedenen Beispielen s.: Lutkevich, CustomGPTs: Example and how to build, TechtTarget, https://www.techtarget.com/whatis/feature/Custom-GPTs-Examples-and-how-to-build (abgerufen 13.12.2024).

[31] Der ITM-GPT kann hier abgerufen und getestet werden: https://www.itm.nrw/ki-tools-fuer-lehre-und-forschung/ (abgerufen 16.4.2025).

plexität des IT-Vertragsrechts zu bewerten, das selbst für erfahrene Praktiker aufgrund der Vielschichtigkeit möglicher Projektkonstellationen eine erhebliche Herausforderung darstellt.

a) Anforderungsprofil und Umsetzung

Die praktische Umsetzung offenbarte eine nicht unwesentliche Diskrepanz zwischen den anvisierten Funktionalitäten und den technischen Möglichkeiten der Custom-GPTs. Zentral sind dabei die Instructions, welche umfassend vorgeben, welche Aufgaben der GPT genau erfüllen soll, wie er dabei vorzugehen hat und welche Aktionen er nicht ausführen soll.

Dem ITM-GPT wurde beispielsweise aufgetragen, stets darauf hinzuweisen, dass die bereitgestellten Informationen und Verträge keine individuelle Rechtsberatung ersetzen können und gerade bei Verträgen eine Prüfung durch einen Rechtsanwalt vorzunehmen ist.

Daneben musste erreicht werden, dass laienhafte Anfragen adäquat in den juristischen Kontext übersetzt werden, sodass der GPT ein zutreffendes Vertragswerk erstellen kann. Durch umfassende Instruktionen, unzählige Tests und Anpassungen konnte erreicht werden, dass der GPT zumeist solange Rückfragen stellt, bis dieser über die für die Erstellung eines Vertrags notwendigen Informationen verfügt. Zumeist gelingt es dem GPT durch die entsprechenden Instruktionen, zentrale Aspekte wie Leistungsumfang, Vergütungsmodelle, Haftungskonstellationen oder Nutzungsrechte in einer auch für Laien zugänglichen Sprache im Wege einer Konversation zu erfassen. So fragt der Bot beispielsweise nicht, ob ein einfacheres oder ausschließliches Nutzungsrecht vereinbart werden soll, sondern versucht den Parteiwillen mit Fragen nach den Modalitäten der Leistungserbringung zu ergründen und anschließend vertraglich umzusetzen.

Für die eigentlichen Erstellung der Verträge hat der ITM-GPT Zugriff auf eine Wissensbasis verschiedener Beispielsverträge, welche aus dem umfangreichen Vertragsportfolio des ITM entstammt. Durch die jahrzehntelange Forschungstätigkeit des ITM im Bereich des IT-Rechts konnten wir dem Bot eine vielschichtige Wissensbasis zur Verfügung stellen, welche auch exotische Konstellationen berücksichtigt. Allgemeines Wissen zum IT-Recht wurde durch die Implementierung des Open-Access-Skriptes IT-Vertragsrecht von Hoeren/Pinelli bereitgestellt.

b) Limitationen

In umfangreichen Tests in Zusammenarbeit mit Praktikern, Studenten und IT-Unternehmen haben sich trotz laufender Anpassungen des Systems verschiedene Grenzen des ITM-GPTs gezeigt. Hervor sticht vor allem eine gewisse Wechselhaftigkeit in der Qualität der generierten Verträge beziehungsweise im Antwortverhalten generell. Der Grund dessen lässt sich indes nicht mit Sicherheit feststellen. Tests lassen aber darauf schließen, dass insbesondere die vorgenannten Instruktionen, die keine nachhaltige Modifikation des Systemverhaltens bewirken, problematisch sind. Die fehlende dauerhafte Integration juristischer Expertise und Analysefähigkeit in Form einer Anpassung des neuronalen Netzwerkes beziehungsweise der Modellparameter scheint gerade in komplexen Anwendungsszenarien zu Schwierigkeit zu führen. So werden die Instruktionen lediglich als temporärer Kontext der jeweiligen Nutzerinteraktion vorangestellt. Damit erfolgt bei jeder Nutzeranfrage eine erneute Interpretation der Instruktionen durch den GPT, wobei die Interpretationstiefe und -qualität in Abhängigkeit von der konkreten Anfragekonstellation erheblich variieren kann. Dieses Problem zeigt sich insbesondere im Rahmen der Vertragsgestaltung, wo die Qualität der generierten Dokumente nicht nur von der Präzession der Nutzeranfrage, sondern auch von der kontextabhängigen Interpretation der juristischen Instruktionen abhängt.

Ähnlicher Gestalt ist auch das zweite Kernproblem. Wesentlich für die Erstellung qualitativer Verträge ist der Rückgriff auf die hinterlegten Vertragsmuster. Während dem System in einigen Fällen eine adäquate Selektion und Integration der relevanten Musterklauseln gelingt, zeigen sich in anderen, i. d. R. komplexeren Fällen, erhebliche Abweichungen von der etablierten Vertragspraxis. Die teilweise erheblich von typisierten Vertragsmustern abweichenden Dokumente adressieren hier zwar oft die Anfrage des Nutzers, setzen diese aber nicht korrekt in etablierter Rechtssprache unter Nutzung gängiger Klauseln um. Während es dem GPT in vielen Fällen beispielsweise ordnungsgemäß gelingt eine Nutzeranfrage nach einem „so weit wie möglich gehenden Haftungsausschluss" rechtssicher in eine standardisierte Haftungsklausel umzusetzen, wird die Haftung in anderen Fällen in rechtswidriger Weise vollständig ausgeschlossen.

Bemerkenswert ist in diesem Zusammenhang auch der erhebliche Kontextualisierungsaufwand, der für eine zielgerichtete Nutzung der Wissensbasis erforderlich ist. Das bloße Hochladen verschiedener Vertragsmuster ist nicht ausreichend. Vielmehr bedarf es einer Strukturierung mittels Inhaltsverzeichnisse sowie spezifischer Anwendungsinstruktionen innerhalb der Vertragsdokumente, die den Verwendungskontext der jeweiligen Vertragsmuster verdeutlichen. Damit konnte die Qualität der generierten Dokumente erheblich erhöht, wenngleich nicht vollständig gewährleistet werden.

c) Praktischer Nutzen niederschwellig individualisierbarer Chat-KIs

Obschon der ITM-GPT den ambitionierten Zielkatalog nicht vollständig erfüllen kann, konnte das Experiment dennoch Erkenntnisse für mögliche Einsatzfelder in der Rechtspraxis liefern.

Im Kontrast zu den vorbeschriebenen Problemen bei der zuverlässigen Erstellung individueller IT-Verträge, gelang die Beantwortung spezialisierter Fragen zum deutschen IT-Recht hingegen in bemerkenswerter Weise. Ein Zugriff auf die im hinterlegten Skriptum zum IT-Recht enthaltenen Informationen erfolgt deutlich kontinuierlicher und der GPT kann relevante Informationen, Rspr. und Erwägungen aus dem Text oder den Fußnoten des Skriptes entnehmen. Ein vielversprechendes Anwendungsfeld ist damit der Aufbau einer spezialisierten Wissensbasis, auf deren Grundlagen komplexe Fragen effizient unter Nennung einschlägiger Rspr. beantwortet werden können (bspw. Urteilsdatenbank).

Die Integration einer Wissensbasis kann auch zur Prüfung von Unterlassungserklärungen genutzt werden. Nach dem Erstellen einer umfassenden Datenbank bereits abgegebener Unterlassungserklärungen kann das System bei der Evaluation neuer Werbeaussagen oder ähnlicher Sachverhalte unterstützend eingesetzt werden und potenzielle Verstöße gegen bestehende Unterlassungsverpflichtungen identifizieren. Im Bereich der Textgenerierung kann der GPT insbesondere als Ideengeber fungieren und bei der Entwicklung initialer Formulierungsvorschläge helfen. In einfacheren Sachverhalten lassen sich auch vollständige Verträge generieren.

Diese Anwendungsszenarien sind dabei stets vor dem Hintergrund der systemischen Limitationen von LLMs zu bewerten. Der Einsatz sollte mithin lediglich unterstützend erfolgen, wobei die finale rechtliche Bewertung beim menschlichen Rechtsanwender verbleibt. Für rechtliche Laien eignen sich diese Systeme damit weniger.

IV. Halluzinatorische Präzision moderner Sprachmodelle

Der Darstellung technischer und implementierungsbezogener Aspekte (individualisierbarer) LLMs folgend, bedarf zuletzt die den Sprachmodellen immanente Tendenz zur Generierung halluzinierter Inhalte einer Betrachtung.

1. Kernproblem

Wie gezeigt basieren moderne LLMs auf probabilistischen Berechnungen. Statt über einen ständigen Zugriff auf eine strukturierte Faktendatenbank, generieren LLMs ihre Aussagen mittels der Vorhersage des jeweils wahrscheinlichsten nächsten Textabschnitts, wobei sie auf während des Trainings erlernte Muster rekurrieren. Der schiere Umfang des Trainingskorpus – beim inzwischen veralteten GPT-3 Modell mehr als 300 Mrd. Wörter und ca. 175 Mrd. Trainingsparameter[32] – führt zu den erstaunlichen Leistungen moderner LLMs. Dieser Umfang übersteigt dabei bei weitem dasjenige, das mit einer deterministischen Informationsspeicherung und -abruf möglich wäre. Der Rückgriff auf Muster, statt einem Zugriff auf faktisches Wissen, kann in Verbindung mit probabilistischer Sprachgenerierung allerdings zu Halluzinationen und sog. überanpassungsartigen Erinnerungen (Overfitting) führen.[33] Letzteres meint eine übermäßig präzise Verinnerlichung bestimmter Trainingsmuster, wodurch es paradoxerweise zu einer Generierung scheinbar präziser (also „fast" richtiger), tatsächlich jedoch faktisch inkorrekter Informationen kommen kann.[34] Besonders problematisch ist hierbei die Tendenz verschiedener Sprachmodelle, Quellenangaben zu generieren, die zwar im jeweiligen Kontext plausibel erscheinen mögen, beim genaueren Hinsehen jedoch vollständig erfunden sind.[35]

Fragt man beispielsweise ChatGPT 4o nach einer juristischen Ausarbeitung zur Abgrenzung der Auftragsverarbeitung gegenüber gemeinsamer Verantwortlichkeit unter Nennung gängiger juristischer Quellen, wird ein inhaltlich und sprachlich weitestgehend zutreffender Text generiert. Die genannten Quellen sind hingegen nur dem ersten Anschein nach richtig. So wird beispielsweise „Gola in: Gola/Heckmann, DS-GVO, 2022, Art. 4 Rn. 44 ff." als Quelle im Rahmen der Begriffsbestimmungen referenziert. Ein Großteil dieser Quelle ist zutreffend (Kommentar, Auflage, Autoren, Herausgeber, kommentierter Artikel), die genaue Fundstelle ist hingegen falsch. Rn. 44 ff. setzen sich inhaltlich mit dem Profiling (Art. 4 Nr. 4 DS-GVO) auseinander. Ähnliches zeigt sich auch bei Urteilen. Im gleichen Text referenziert ChatGPT 4o richtigerweise das „EuGH, Urteil vom 29.07.2019, C-40/17 – Fashion-ID", verweist dabei aber auf Rn. 64, welche lediglich eine vorgelegte Frage wiedergibt, nicht aber die Aussage im Text belegt. Solch fehlerhafte Generierungen vermeintlich präziser Quellenangaben (oder anderer Halluzinationen) beruhen auf der probabilistischen Synthese sprachlicher Muster, die im Trainingsprozesses extrahiert und bestimmten Kontexten zugeordnet wurden. „Kontext" ist hierbei allerdings nicht auf einen semantischen oder fachlichen Rahmen bezogen, sondern meint vielmehr die statistische Nähe zwischen Token innerhalb des transformergestützten Vektorraumes.[36] Bei den oben genannten Beispielen lässt sich vermuten, dass das Sprachmodell spezifische Elemente (dh Token) dieser Quellenangaben (bspw. Kommentatoren, Titel, Urteilsdaten etc) in unterschiedlichen Kombinationen in ähnlichen Kontexten während des Trainings gesehen und daraus Muster entwickelt hat. Die spätere Rekonstruktion solcher Quellen bei der Erstellung eines Textes erfolgt dann gerade nicht durch einen gezielten Abruf gespeicherten Wissens, sondern durch die oben beschriebene, probabilistische Generierung plausibler Kombinationen auf Basis mus-

---

[32] Singh et. al., CODEFUSION: A Pre-trained Diffusion Model for Code Generation, 2023 Conference on Empirical Methods in Natural Language Processing, S. 11700.

[33] Rawte/Sheth/Das, A Survey of Hallucination in „Large" Foundation Models, https://doi.org/10.48550/arXiv.2309.05922.

[34] Lesenswert hierzu: Hense MMR 2024, 449 (450).

[35] Vertiefend zur Zuverlässigkeit von LLMs bei der Erstellung wissenschaftlicher Quellenangaben: Mugaanyi et. al. J Med Internet Res. Vol. 26 2024, https://doi.org/10.2196/52935.

[36] Die semantische und syntaktische Nähe verschiedener Begriffe (bspw. „Athen" zu „Griechenland" oder „möglich" zu „unmöglich") innerhalb eines mathematischen Vektorraumes anschaulich beschreibend: Mikolov et. al., Efficient Estimation of Word Representations in Vector Space, https://doi.org/10.48550/arXiv.1301.3781, S. 5 f.

terabhängiger, statistischer Wahrscheinlichkeit. Während Autorennamen o. Ä. im Rahmen eines statistischen Kontextes relativ gleichbleibend und daher in der Ausgabe korrekt sind, wird das Sprachmodell während des Trainings eine Vielzahl verschiedener Rn.-Angaben im ähnlichen Kontext inkorporiert haben, wodurch die probabilistische Generierung mangels eindeutig wahrscheinlicher Antwort fehlgeht. Fehlt dem genutzten Sprachmodell zudem eine interne Limitierung, welche zur Verhinderung von Halluzinationen gewisse Anfragen (teilweise) verweigert, wird das Sprachmodell ggf. auch ganze Urteile oder wörtliche Zitate erfinden und mit falschen Quellenangaben versehen.[37]

2. Mitigationsmaßnahmen

Den Risiken halluzinierter Inhalte kann durch verschiedene Maßnahmen begegnet werden.

Zuvörderst ist ein systematisches Prompt-Engineering wichtig, welches durch präzise Formulierungen und strukturierte Mehrschritt-Anfragen die Wahrscheinlichkeit fehlerhafter Aussagen reduzieren kann. Negativ-Prompts, die das Modell explizit zur Vermeidung von Spekulationen oder nicht verifizierbaren Quellen auffordern, können die Präzision der Ausgabe weiter erhöhen. Eine weitere Verbesserung der Ausgabequalität lässt sich zudem durch die Integration zusätzlichen Wissens, insbesondere bei Zugriffsmöglichkeiten auf externe Datenbanken erreichen. Dies ermöglicht Retrieval-Augmented-Generation (RAG), bei dem die Sprachmodelle auf strukturierte Faktendatenbanken zugreifen und deren Inhalte in die Textgenerierung einbinden.[38] In einem reduzierteren Maße ist dies auch durch den Nutzer möglich, indem dieser eigene Quellen oder Informationen zur Verfügung stellt. Damit ist eine verlässliche Verifizierung von Quellenangaben möglich und die Wahrscheinlichkeit halluzinierter Inhalte wird maßgeblich reduziert (bspw. Urteile im Volltext). Ein weiterer effektiver Ansatz zur Vermeidung erfundener Angaben besteht in der Nutzung dedizierter Recherche-KIs, welche mittels Datenbank- oder Internetsuchmaschinenanbindungen verlässlich nach den gewünschten Informationen suchen können und die entsprechenden Quellen zumeist direkt verlinken.[39]

Schlussendlich entscheidend ist aber die Sensibilisierung der Nutzer für die systemimmanenten Limitationen auch modernster Sprachmodelle und gerade in sensiblen Anwendungsbereichen wie der Rechtswissenschaft ist eine menschliche Verifizierung unerlässlich.

V. Fazit & Schlusswort

Die von Spindler bereits lange prognostizierte rasante Entwicklung künstlichen Intelligenz, zeigt sich heute besonders im Bereich der Large Language Models. Die hier vorgestellte Fallstudie zum ITM-GPT soll dabei exemplarisch aufzeigen, wie sich Rechtsanwender unterschiedliche Systeme zu Nutzen machen können. Diese Entwicklungen induzieren aber auch eine fundamentale Veränderung der rechtlichen Arbeitsweise, die einer aktiven Gestaltung durch Wissenschaft und Praxis gleichermaßen bedürfen. Bei der Nutzung auch fortschrittlicher LLMs sollte sich der Nutzer stets bewusst sein, dass die Modelle ledig-

---

[37] So einem US-Anwalt geschehen, der ein halluziniertes Urteil vor Gericht als Argumentationsgrundlage nutzte, Bohannon, Lawyer Used ChatGPT In Court – And Cited Fake Cases, Forbes, https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/ (abgerufen 13.12.2024).

[38] Ausf.: S. Hofstätter et. al., Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 1437–1447.

[39] Beispiele solcher Systeme: https://elicit.com/, https://consensus.app, https://www.perplexity.ai/ oder ScholarGPT; für einen kurzen Überbl.: https://law.mpg.de/news/ai-tools-for-legal-research/ (abgerufen 13.12.2024).

lich als eine hochentwickelte Repräsentation statistischer Zusammenhänge agieren, die zwar scheinbar kohärente und oftmals zutreffende Texte genieren können, dabei aber keine inhaltliche Faktizität garantieren oder über ein tatsächliches Verständnis im Sinne eines juristischen oder technischen Begriffsrahmens aufweisen.[40] Nicht zuletzt diese strukturelle Begrenzung macht die Nutzung moderner LLMs – auch wenn sie für einen bestimmten Anwendungsfall individualisiert wurden – zwar als unterstützendes Werkzeug interessant, erfordert aber auch stets eine kritische Überprüfung der generierten Inhalte.

Die von Spindler stets angemahnte Verbindung von technischem Verständnis und juristischer Expertise wird dabei wichtiger denn je – nicht nur für die weitere wissenschaftliche Auseinandersetzung mit KI-Systemen, sondern auch für deren praktische Nutzung im juristischen Alltag. Einmal bei einem Workshop über die rechtlichen Herausforderungen der Digitalisierung wurde Spindler von einem Teilnehmer gefragt, wie er es schaffe, immer auf dem neuesten Stand der Technologien und Gesetzgebung zu bleiben. Mit einem schelmischen Lächeln antwortete er: „Das Geheimnis ist simpel: Ich frage meine Studierenden. Sie kennen die Technik, und ich mache mir dann die juristischen Sorgen." Wer Spindler kannte, wusste, dass er damit auf charmante Weise seinen Teamgeist und seine Fähigkeit zur Zusammenarbeit hervorhob – und hoffentlich auch seine Freude an unserem eher technischen Ausweg auf die Möglichkeiten und Grenzen juristischer Chatbots gehabt hätte.

[40] Schlussendlich sind moderne LLMs nichts anderes als „stochastische Papageien". Dieser Begriff wurde geprägt durch Bender et al., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, https://doi.org/10.1145/3442188.3445922.