Wolfgang Eppler, Reinhard Heil

Generative KI für TA

Viele Wissenschaftler*innen nutzen generative KI in ihrer wissenschaftlichen Arbeit. In der Technikfolgenabschätzung (TA) arbeitende Menschen sind hierbei nicht ausgenommen. Der Umgang der TA mit generativer KI ist ein doppelter: Einerseits wird generative KI für die TA-Arbeit genutzt, andererseits ist generative KI Gegenstand der TA-Forschung. Der folgende Beitrag geht, nachdem er kurz das Phänomen generativer KI umrissen und Anforderungen für deren Einsatz in der TA formuliert hat, ausführlich auf die strukturellen Ursachen der mit ihr verbundenen Probleme ein. Generative KI wird zwar ständig weiterentwickelt, die strukturell bedingten Risiken bleiben jedoch bestehen. Der Artikel schließt mit Lösungsvorschlägen und kurzen Hinweisen zu deren Umsetzbarkeit sowie einigen Beispielen für den Einsatz generativer KI in der TA-Arbeit ab.

1. Phänomene der generativen KI

Ein herausragendes Merkmal generativer KI (im Folgenden synonym verwendet mit Chatbot und multimodalem Sprachmodell LLM) ist, dass sie es ermöglicht, mit Maschinen schriftlich oder mündlich in natürlicher Sprache zu kommunizieren. Chatbots, wie ChatGPT, Gemini, LLama oder Claude, die auf sogenannten großen Sprachmodellen basieren, lassen sich sehr einfach, ohne Einarbeitung nutzen. Man kann mit ihnen plaudern (Tuschling et al. 2023), ihnen Fragen stellen, sie können Texte zusammenfassen und aus Eingaben, sogenannten Prompts, Texte und Bilder in beliebigen Formen und Stilen erzeugen (Yin et al. 2024). Die Anwendungen sind aber nicht auf natürliche Sprache beschränkt, sie können unter anderem auch Programmcode erzeugen. Mittlerweile führen Chatbots (als Agenten, beispielsweise durch Large Action Models (Hille 2025)) Aktionen aus, sie können im Internet suchen oder Unternehmensdaten verarbeiten (Gao et al. 2024).

Die LLMs sind jedoch an erster Stelle Sprachmodelle und keine Wissensmodelle. Sie geben nicht ein einmal gelerntes Wissen wieder, sondern generieren, anhand der während des Trainings erkannten statistischen Zusammenhänge, neue Texte, die faktisch richtig sein können, aber nicht müssen. Sie können Fehlinformationen (*misinformation, disinformation, hallucinations* (Deng et al. 2024))

erzeugen, was ihre Vertrauenswürdigkeit untergräbt. Da die Texte, mit denen Sprachmodelle trainiert werden, alle Arten von destruktiven, hasserfüllten und toxischen Aussagen enthalten, finden sich Vorurteile, Voreingenommenheit, verzerrte Ansichten (*bias*) auch in der Ausgabe generativer KI wieder, oft in verstärktem Maße (Parrish et al. 2022). Um das Erzeugen solcher unerwünschten, teils von Nutzenden gezielt angeregten Ausgaben (Fehlausrichtung (*misalignment*)) zu verhindern, werden Korrekturmaßnahmen durchgeführt. Dieses sogenannte *alignment* bleibt jedoch nicht ohne Folgen, es kann selbst zu Verzerrungen führen (Askell et al. 2021). Generative KI ist intransparent, es ist zumeist nicht möglich, genau nachzuvollziehen, wie eine Ausgabe zustande kommt. Es gibt zwar eine große Zahl von Forschungsansätzen, die versuchen, KI erklärbar zu machen, diese können aber bisher die in sie gesetzten Erwartungen nicht erfüllen (zum Beispiel Wu et al. 2024; Renftle et al. 2022).

2. Strukturelle Ursachen von Problemen derzeitiger KI

Eine paradigmatische Aufgabe der TA ist die Politikberatung (Grunwald 2023). Sie erfordert verlässliche, kluge, situationsgerechte Lösungen. Um dies zu gewährleisten, müssen die von generativer KI erzeugten Ausgaben verständlich, nachvollziehbar, kontrollierbar, kohärent, diskriminierungsfrei und erklärbar sein sowie Quellen und verwendete Daten offengelegt werden.

Im Folgenden werden die strukturellen Ursachen von Problemen derzeitiger KI untersucht, um besser einschätzen zu können, wo generative KI in der TA eingesetzt werden kann und wo nicht. Es lassen sich zumindest acht strukturelle Ursachen derzeitiger KI-Probleme aufweisen.

(1) Datenqualität

Der größte Teil der für das Training der KI verwendeten Daten kommt aus dem Internet (siehe Liu et al. 2024b). Die Qualität dieser Daten ist oft schlecht, da unter anderem viele synthetische Daten (von Maschinen erzeugt, Systemprotokolle, …) enthalten sind (Shumailov et al. 2024), Daten mehrfach vorkommen (Carlini et al. 2023; Deng et al. 2024) oder, gerade bei neuen gesellschaftlichen Entwicklungen, widersprüchlich sind (Augenstein et al. 2024). Zudem ist es praktisch unmöglich Datensätze dieser Größenordnung auf sachliche Richtigkeit zu prüfen. Neben den Internetdaten werden zum Training auch digitalisierte Bücher und Zeitschriften genutzt. Die Daten sind auf keinen Fall gleichverteilt, sie ent-

halten Lücken, die dem Anwender verborgen bleiben. Verteilungsverschiebungen[1] führen mit dazu, dass Modelle in neuen Situationen falsch generalisieren (Liu et al. 2024a; Vafa et al. 2024). Die Anbieter generativer Modelle bereinigen fast immer die Originaldaten und setzen Clickworker zum Annotieren ein. Wie dies genau geschieht und welche Anpassungen vorgenommen werden, bleibt meist im Dunkeln (Rohde 2024; Albalak et al. 2024).

(2) Fehlausrichtung

Zwar verfügen auf generative KI basierende Chatbots über beeindruckende sprachliche Fähigkeiten, sie geben aber immer wieder ethisch zweifelhafte Antworten. Um dem zu begegnen, wurden verschiedene, aufeinander aufbauende Fehlausrichtungskorrekturen (*alignment*) entwickelt. Als erstes ist das Fine-Tuning zu nennen. Fine-Tuning bedeutet, dass das Training anhand großer Datenmengen (Pre-Training) ergänzt wird, indem das Sprachmodell mit annotierten Anfrage-Antwortpaaren trainiert wird (Ouyang et al. 2022). Diese Anfrage-Antwortpaare sind im Vergleich zu den Trainingstexten des Pre-Trainings anwendungsbezogener, von Menschen explizit bewertet (*supervised*) und von deutlich geringerer Anzahl als die Texte des Pre-Trainings. Ziel ist es, toxische Inhalte, also Diskriminierungen, Hassreden, destruktive Handlungen, und so weiter zu erkennen, abzumildern oder zu korrigieren. Da die zusätzlichen Trainingsdaten stärker auf die alltägliche Nutzung der Chatbots zugeschnitten sind und für entsprechende Anwendungen die Datenlage verbessern, sollen sie zugleich die Anzahl der Fehlinformationen reduzieren. Ähnlich wie das Fine-Tuning dient das *Instruction Tuning* dazu, die Modellantworten auf Alltagsanfragen zu verbessern. Eine ganz andere Art von Ausrichtung ist durch die Verwendung von Prompterweiterungen möglich. Der Systemprompt teilt dem generativen Modell vor jeder Benutzeranfrage bereits einige Informationen mit, beispielsweise, wie das Modell heißt, aber auch, wie das Modell in bestimmten Situationen reagieren soll. Beispielsweise wird dem System so mitgeteilt, dass es die Beantwortung einer Aufgabe Schritt für Schritt angehen soll (*Chain of Thought (CoT) Prompt* (Wei et al. 2023)), um so seine eigenen Ausgaben in den Lösungsweg mit einzubeziehen. Fine-Tuning und Prompts können von den Nutzenden nochmals zusätzlich gestaltet werden, um das generative Modell noch stärker auf ihre Anforderungen zuzuschneiden.

1 Die Menge der Trainingsdaten ist nicht gleichverteilt. Sie weist an verschiedenen Stellen des Datenraums Häufungen beziehungsweise Löcher auf. Wird ein Modell häufig in einem Bereich angewandt, der von den Daten schlecht abgedeckt ist (Datenloch), weisen die Anwendungsdaten gegenüber den Trainingsdaten eine verschobene Verteilung auf.

Diese Alignment-Methoden haben sehr zum Erfolg der generativen KI beigetragen. Sie haben aber die Zuverlässigkeit der Modelle nicht genügend erhöht, sie können sie teilweise sogar verringern. So können durch Fine-Tuning Inhalte, die durch Pre-Training gelernt wurden, überschrieben und vergessen werden (*catastrophic forgetting* (Ven et al. 2024)), sich toxische Inhalte einschleichen oder eine ausgewogene Datenverteilung aus dem Gleichgewicht gebracht werden (Qi et al. 2023). Woran liegt diese Resistenz gegenüber allen Korrekturmaßnahmen? Ein Grund ist, dass Menschen und ML-Modelle oft nicht die gleichen Merkmale nutzen, um Kategorien zu bilden (*Proxy-Effect* (John et al. 2024)). So kommt es vor, dass eine KI generalisierende Merkmale herausbildet, die nur aufgrund von Randeffekten, wie zum Beispiel ein Wasserzeichen, eine Unschärfe im Bild oder ein belanglos eingeflochtener Nebensatz, die für einen Menschen unerheblich sind, zum gewünschten Ergebnis führen. Dies führt auch dazu, dass KI unerwartete oder andere Fehler macht als Menschen.

(3) Kontext-Inhalt-Herausforderung

Es gibt ein Spannungsverhältnis zwischen der Benutzereingabe (Kontext) und dem, was ein Sprachmodell während des Trainings gelernt hat, da die Benutzereingabe Aussagen enthalten kann, die dem gelernten Inhalt widersprechen. Die spannende Frage ist dann, welcher Anteil sich durchsetzt (Min et al. 2022; Dai et al. 2023; Kossen et al. 2024). Bekannt ist, dass das Anführen von Beispielen im Eingabe-Kontext, sogenanntes In-Context Learning oder Few Shot Prompting, die Ausgabequalität eines Sprachmodells stärker verbessern kann als das Fine-Tuning (Brown et al., 2020). Das ist besonders bei Aufgaben einsichtig, die eine Ausgabe in Tabellenform erlauben und als annotierte Eingabe die drei oder vier ersten Zeilen der Tabelle explizit mitgeteilt bekommen. Für den Rest der Tabelle wird nur noch die Eingabe vorgegeben und die Ausgabe vom Sprachmodell ergänzt. Diese Vorgehensweise suggeriert, dass von außen die Form vorgegeben wird, während der Inhalt durch Informationen des Modells bestimmt wird. Es zeigt sich jedoch auch, besonders bei umfangreicherem Kontext, dass auch Inhalte des Kontextes in die Antwort übernommen werden – unabhängig davon, ob sie richtig oder falsch sind (sycophancy (deutsch: Kriecherei)). Die Intransparenz der Daten wie auch die Komplexität der Modellarchitektur sorgen dafür, dass nicht vorhergesagt werden kann, in welchen Fällen sich der Kontext und wann sich der Inhalt durchsetzt. Das Ergebnis ist unbestimmt.

(4) Reproduktion

Der Trainingsprozess geschieht nicht fortlaufend, nicht während der Dialoge mit den Chatbots, sondern wird einmalig mit riesigen Datensätzen durchgeführt und kann Wochen dauern. Deshalb sind große Sprachmodelle außerstande, eigene Erfahrungen zu machen. Sie können nicht kontinuierlich lernen. An Trainingsalgorithmen, die dies ermöglichen, wird intensiv geforscht (Shi et al. 2024)

Aktuelle gesellschaftliche Veränderungen und menschliche Erfahrungen sickern so erst allmählich in Texte ein, mit denen dann eine KI trainiert wird. Die KI hinkt also der gesellschaftlich-technischen Entwicklung im Vergleich zum Menschen hinterher. So ergibt sich eine zeitliche Kluft zwischen einer aktuellen Benutzeranfrage und dem Aufkommen einer Veränderung mit den nacheinander folgenden Schritten:

- gesellschaftliche Veränderungen, die noch nicht in Sprache gefasst wurden.
- textliche Verarbeitung eines Themas in der Gesellschaft (Informationsquelle)
- Festlegung des Korpus mit Trainingsdaten
- Release-Zeitpunkt des Sprachmodells
- Benutzeranfrage an eine KI

Die KI braucht Texte schreibende Menschen für ihre Reproduktion. Synthetische Texte können diese Grundlage nicht liefern. Das von (Harnad 1990) zuerst beschriebene Grounding-Problem für symbolische KI ist von Bender und Koller (2020) für die generative KI bestätigt worden. Unter anderem die Arbeiten der Philosophen Habermas (1981) und Brandom (1994) zeigen, dass für einen Realitätsbezug kontinuierliches Lernen und verantwortungsbewusste Setzungen unabdingbar sind. Dies können Sprachmodelle bisher nicht leisten.

(5) Fehlen sozialer Perspektiven

Von einer KI erwarten wir, nachdem wir eine falsche Antwort korrigiert haben, dass das Modell die Korrektur annimmt und das nächste Mal berücksichtigt. Unter (4) Reproduktion konnten wir sehen, dass dazu kontinuierliches Lernen notwendig wäre, wozu heutige Modelle nicht in der Lage sind. Eine andere Situation stellt sich ein, wenn bei der Bewertung einer Handlung die Meinungen auseinander gehen. Die Klärung erfolgt dann nicht mehr direktiv wie in einem asymmetrischen Eltern-Kind- oder Lehrer-Schüler-Verhältnis, sondern symmetrisch durch einen Diskurs, in dem über den Austausch von Argumenten geklärt wird, wer Recht hat. Um einen solchen Diskurs über Wahrheit zu führen, muss man in der Lage sein, verschiedene soziale Perspektiven einnehmen zu können

(Habermas 1981; Brandom 1994). Die Perspektive generativer KI ist funktional. Ihr fehlt die normative Komponente, die einen menschlichen Gesprächspartner auszeichnet. Warum setzt ein Wahrheitsdiskurs solche sozialen Perspektiven voraus? Eine zweite Perspektive generalisiert die Meinung einer einzigen Perspektive, die dadurch intersubjektiviert wird. Fehlt die zweite Perspektive, fehlt ein normatives Korrelat, das die richtige Verwendung eines Begriffs bestätigt. Die soziale Perspektive ermöglicht es, nicht nur zwischen bloßen Behauptungen und begründbaren Aussagen zu unterscheiden, sondern die Sprecher übernehmen zudem Verantwortung für das von ihnen Gesagte, sie gehen Verpflichtungen ein.

(6) Weltmodell

Wenn im Rahmen der generativen KI von Weltmodell gesprochen wird, ist damit zumeist ein physikalisches Weltmodell mit körperlichen Fähigkeiten in einer raumzeitlichen Welt mit physikalischen Gesetzmäßigkeiten gemeint (Xu et al. 2024). Die körperlichen Fähigkeiten sind im Gegensatz zu den in den Sprachmodellen zweifellos vorhandenen sprachlichen Fähigkeiten zu sehen. Demgegenüber gibt es auch phänomenologische Weltmodelle, welche eine Welt mit Erfahrungen und Wahrnehmungen abbilden. Dieser Aspekt wurde unter (4) Reproduktion behandelt. Psychologische Weltmodelle umfassen die mentale Welt mit unserem Alltagsverstand und Schlussfolgern (siehe unten (7) Reasoning). Bei allen drei verschiedenen Arten von Weltmodellen spielen Veränderungen eine Rolle, in Form von Bewegungen (physikalisch), Erfahrungen (phänomenologisch) oder Lernen (psychologisch).

Die zurzeit diskutierten Weltmodelle für die KI lassen sich in drei aufeinanderfolgenden Stufen einteilen: Videogeneratoren, Simulationsmodelle und reale Weltmodelle. Ein vielbeachteter Videogenerator ist SORA von OpenAI (Brooks et al. 2024). Durch Auswertung vieler Videos gelingt es, die in Videos beobachteten Bewegungen zu generalisieren – allerdings ohne die dahinterliegenden physikalischen Gesetze erfahren zu haben. In vielen mit SORA erzeugten Videos können Unstimmigkeiten beobachtet werden wie eine fehlende Objektpersistenz und physikalische Inkohärenzen über längere Zeit. Simulationsmodelle kommen in Computerspielen vor, in denen Agenten in einer künstlichen Welt erfundenen Gesetzen unterworfen sind. Damit wirkt die künstliche Welt auf die Agenten zurück, was bei Videogeneratoren fehlt. Allerdings fehlen der künstlichen Welt die Unvorhersehbarkeiten einer realen Welt. Dies erfahren Roboter, die vorher in Simulationen trainiert wurden und anschließend der realen Welt ausgesetzt

werden. Es gibt zurzeit aber keine Trainingsalgorithmen aus der generativen KI, die das leisten (siehe (4) Reproduktion mit fehlendem kontinuierlichem Lernen).

Fehlt ein realistisches Weltmodell, wird eine generative KI immer wieder inkohärente Ausgaben erzeugen und mit raumzeitlichen Zuordnungen Schwierigkeiten haben.

(7) Reasoning

Reasoning (Schlussfolgern mit gesundem Menschenverstand) besitzt ein weites Bedeutungsspektrum. Sun et al. (2023) (siehe Abbildung 1) zeigen, wie vielfältig dieser Begriff gebraucht wird. Generative KI schneidet, was verlässliches Schlussfolgern angeht, noch ziemlich schlecht ab. Als Gründe für den fehlenden Alltagsverstand können unter anderem ein fehlendes physikalisches Weltmodell (siehe *(6) Weltmodell*) und fehlende körperliche und kontinuierliche Erfahrungen genannt werden. Die Reflexion generativer KI mithilfe von Zwischenschritten (Chain of Thought mit "*hidden CoT prompt*", siehe *(2) Fehlausrichtung*) soll eine weitere *soziale Perspektive (5)* implementieren, um so logisches Schließen verlässlicher zu machen. Angeblich sind die großen Sprachmodelle damit auf einem guten Weg, besser zu werden. Ein Durchbruch wurde angeblich mit ChatGPT-4o1 geschafft. Kurz nach Einführung des neuen Systems wurde jedoch bereits aufgezeigt, welche Lücken sich beim zielgerichteten Schlussfolgern zeigen. Das komplizierte und falsche Schließen (zum Beispiel bei einem abgewandelten Flussüberquerungsrätsel) ist erklärbar, wenn man davon ausgeht, dass vergleichbare Aufgaben bereits im Trainingsdatenset vorhanden waren und deren Struktur fälschlich auf neue Aufgaben übertragen wurden (Mirzadeh et al. 2024). Wird die Aufgabe in einem wesentlichen Detail abgewandelt, geht das Schließen fehl.

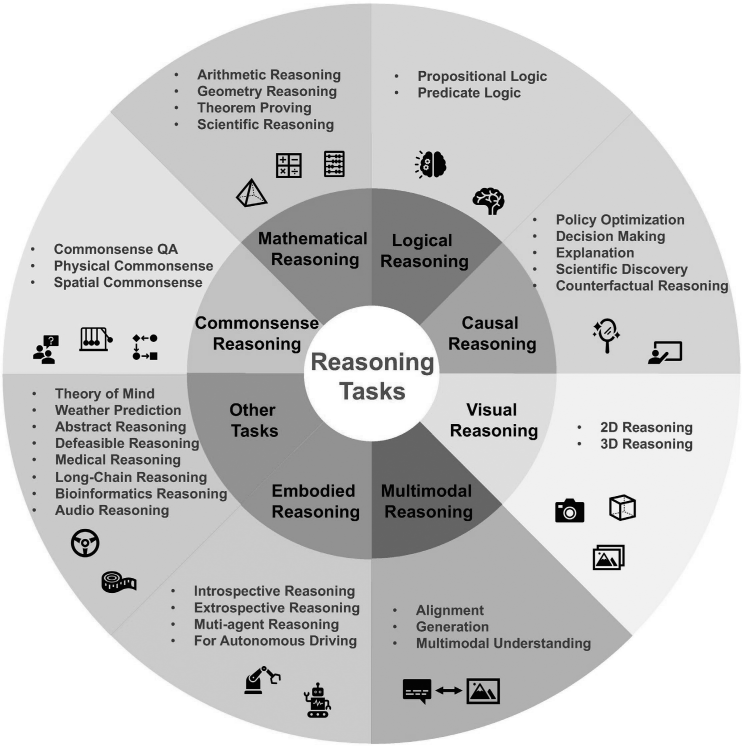

Abbildung 1: Reasoning: Verschiedene Arten von Schlussfolgern. Quelle: Sun et al. 2023

3. Einsatzmöglichkeiten generativer KI für die TA

Aufgrund der angeführten Schwächen ist es ratsam, generative KI sehr bedacht und nur dann einzusetzen, wenn man in der Lage ist, die Ergebnisse zu überprüfen. Für die Technikfolgenabschätzung ist die KI sowohl Untersuchungsgegenstand als auch mögliches Werkzeug. Im Folgenden werden, ohne Anspruch auf Vollständigkeit, in der TA anfallende Aufgaben daraufhin überprüft, inwiefern sie angesichts der zuvor herausgearbeiteten Herausforderungen mit Hilfe generativer KI bearbeitet werden können.

Das sogenannte Horizon-Scanning (Hungerland 2025), wie es unter anderem am Büro für Technikfolgenabschätzung beim Deutschen Bundestag (TAB) durchgeführt wird, eignet sich gut um die Möglichkeit und Grenze des Einsatzes generativer KI in der TA zu verdeutlichen. Am Anfang des Horizon-Scannings, wie jedes TA-Projektes, steht die Informationssammlung mittels einer strukturierten Suche nach und Auswertung von Quellen. Ausgewertet werden Printmedien, Onlinequellen (Webseiten, Blog, Datenbanken et cetera), wissenschaftliche Publikationen, Reports/Berichte, es werden aber auch Expert*innen und Laien befragt. Es gibt mittlerweile eine Unzahl an auf generativer KI basierender KI-Tools, die die Auswertung solcher Quellen erleichtert. Alle diese Tools sind jedoch unzuverlässig. Ob beispielsweise eine automatisch erstellte Literaturübersicht alle relevanten Quellen enthält, sachlich richtig ist, keine Übertreibungen, erfundenen Quellen und anderer sogenannte Halluzinationen enthält muss geprüft werden, ebenso automatisiert erstellte Interviewtranskripte. Generative KI kann zwar das Schreiben von Überblicken erleichtern, aber auf keinen Fall vollständig übernehmen. Zur Informationssammlung gehört auch die Integration und Anreicherung der gefundenen Quelle, qualitative Inhaltsanalysen, sowie Interpretation, Diskussion und Validierung der gefundenen Informationen. Generative KI kann hier insbesondere dazu genutzt werden, um bei der Clusterung von Informationen zu helfen und Anregungen zu geben. Solange entsprechenden Vorschlägen nicht blind vertraut wird, kann dies die Informationsauswertung und -aufbereitung erleichtern. Gleiches gilt für den auf die Informationssammlung folgenden Schritt des Horizon-Scannings, die Ideenaufbereitung. Ihre eigentlichen Vorteile kann die generative KI aber insbesondere bei der sprachlichen und graphischen Aufbereitung der Ergebnisse ausspielen: Erzeugung von Fließtexten aus Aufzählungen und umgekehrt, Erstellung von Folien und graphischen Darstellungen, Textkorrektur und -erzeugung, Übersetzungsunterstützung et cetera. Selbstverständlich müssen auch bei diesem Schritt alle automatisch generierten Ergebnisse auf Korrektheit geprüft werden.

4. Konklusion

Grob zusammengefasst sollte derzeit die Anwendung generativer KI in der TA darauf beschränkt werden, sie als Ideengeber und als Unterstützung zu nutzen. Ungeprüft dürfen die Ergebnisse nicht verwendet werden, man darf sich nicht auf sie verlassen. Damit unterscheidet sich der Einsatz von generativer KI in der TA kaum von ihrem Einsatz in anderen wissenschaftlichen Bereichen.

Wir sollten aber nicht vergessen, dass es in der TA um mehr geht als möglichst schnell möglichst viele Informationen zu sammeln und zusammenzufassen. Wir leiden unter akademischem FOMO (fear of missing out), wir haben ständig Angst, etwas Wichtiges zu übersehen. Generative KI wird oft als Retter in der Not dargestellt, als einzige Möglichkeit, der Informationsflut Herr zu werden. Das mag zwar stimmen, aber die Idee, dass über mehr Informationen zu verfügen immer besser ist, ist von vorneherein falsch. Mehr Informationen tragen oft gar nicht nennenswert zu einem besseren Verstehen bei. Intensives Lesen ist ein wichtiger Teil des Verstehens, ebenso wie das Zusammenfassen und Schreiben. Was uns Maschinen tatsächlich niemals werden abnehmen können, ist zu verstehen und wir müssen darauf achten, dass wir die zum Verstehen notwendigen Fähigkeiten nicht aus Zeitdruck und Bequemlichkeit verlernen.

Literatur

Albalak, A.; Elazar, Y.; Xie, S. M.; Longpre, S.; Lambert, N.; Wang, X.; Muennighoff, N.; Hou, B.; Pan, L.; Jeong, H.; Raffel, C.; Chang, S.; Hashimoto, T.; Wang, W. Y. (2024): A Survey on Data Selection for Language Models (No. arXiv:2402.16827). DOI: 10.48550/arXiv.2402.16827

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Kernion, J.; Ndousse, K.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Kaplan, J. (2021): A General Language Assistant as a Laboratory for Alignment (No. arXiv:2112.00861). DOI: 10.48550/arXiv.2112.00861

Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G. L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; Hovy, E.; Ji, H.; Menczer, F.; Miguez, R.; Nakov, P.; Scheufele, D.; Sharma, S.; Zagni, G. (2024): Factuality challenges in the era of large language models and opportunities for fact-checking. Nature Machine Intelligence 6(8), S. 852–863. DOI: 10.1038/s42256–024–00881-z

Bender, E. M.; Koller, A. (2020): Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, S. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463

Brandom, R. (1994); Making it explicit: Reasoning, representing, and discursive commitment. Harvard University Press

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; Ramesh, A. (2024): Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators [aufgerufen am 11.12.2025]

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, Ilya; Amodei, D. (2020): Language Models are Few-Shot Learners (No. arXiv:2005.14165). DOI:10.48550/arXiv.2005.14165

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; Zhang, C. (2023): Quantifying Memorization Across Neural Language Models (No. arXiv:2202.07646). DOI: 10.48550/arXiv.2202.07646

Dai, D.; Sun, Y.; Dong, L.; Hao, Y.; Ma, S.; Sui, Z.; Wei, F. (2023): Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers. In: Rogers, A.; Boyd-Graber, J.; Okazaki, N. (Hg.) (2023): Findings of the Association for Computational Linguistics: ACL 2023, S. 4005–4019. DOI: 10.18653/v1/2023.findings-acl.247

Deng, C.; Duan, Y.; Jin, X.; Chang, H.; Tian, Y.; Liu, H.; Zou, H. P.; Jin, Y.; Xiao, Y.; Wang, Y.; Wu, S.; Xie, Z.; Gao, K.; He, S.; Zhuang, J.; Cheng, L.; Wang, H. (2024): Deconstructing The Ethics of Large Language Models from Long-standing Issues to New-emerging Dilemmas (No. arXiv:2406.05392). DOI: 10.48550/arXiv.2406.05392

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. (2024): Retrieval-Augmented Generation for Large Language Models: A Survey (No. arXiv:2312.10997). DOI: 10.48550/arXiv.2312.10997

Grunwald, A. (2023): Technikfolgenabschätzung. Baden-Baden

Habermas, J. (1981): Theorie des kommunikativen Handelns. Berlin

Harnad, S. (1990): The symbol grounding problem. In: Physica D: Nonlinear Phenomena 42(1), S. 335–346. DOI: 10.1016/0167–2789(90)90087–6

Hille, M. (2025). Large Action Models – neue KI-basierte Interaktionsformen mit digitalen Technologien. TAB-Themenkurzprofil 78. DOI: 10.5445/IR/1000179508

Hungerland, T. (2025, June 11). Horizon-Scanning (KIT). https://www.tab-beim-bundestag.de/projekte_horizon-scanning.php [aufgerufen am 11.12.2025]

John, Y. J.; Caldwell, L.; McCoy, D. E.; Braganza, O. (2024): Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems. Behavioral and Brain Sciences, 47, e67. DOI: 10.1017/S0140525X23002753

Kossen, J.; Gal, Y.; Rainforth, T. (2024): In-Context Learning Learns Label Relationships but Is Not Conventional Learning (No. arXiv:2307.12375). DOI: 10.48550/arXiv.2307.12375

Liu, B.; Zhan, L.; Lu, Z.; Feng, Y.; Xue, L.; Wu, X.-M. (2024a): How Good Are LLMs at Out-of-Distribution Detection? (No. arXiv:2308.10261). DOI: 10.48550/arXiv.2308.10261

Liu, Y.; Cao, J.; Liu, C.; Ding, K.; Jin, L. (2024b): Datasets for Large Language Models: A Comprehensive Survey (No. arXiv:2402.18041). DOI: 10.48550/arXiv.2402.18041

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; Zettlemoyer, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? (No. arXiv:2202.12837). DOI: 10.48550/arXiv.2202.12837

Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; Farajtabar, M. (2024): GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models (No. arXiv:2410.05229). DOI: 10.48550/arXiv.2410.05229

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; Lowe, R. (2022): Training language models to follow instructions with human feedback (No. arXiv:2203.02155). DOI: 10.48550/arXiv.2203.02155

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; Bowman, S. R. (2022): BBQ: A Hand-Built Bias Benchmark for Question Answering (No. arXiv:2110.08193). DOI: 10.48550/arXiv.2110.08193

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; Henderson, P. (2023): Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! (No. arXiv:2310.03693). DOI: 10.48550/arXiv.2310.03693

Renftle, M.; Trittenbach, H.; Poznic, M.; Heil, R. (2022): Explaining Any ML Model? – On Goals and Capabilities of XAI (No. arXiv:2206.13888). DOI: 10.48550/arXiv.2206.13888

Rohde, F. (2024): Taking (policy) action to enhance the sustainability of AI systems. https://www.ioew.de/publikation/taking_policy_action_to_enhance_the_sustainability_of_ai_systems [aufgerufen am 11.12.2025]

Shi, H.; Xu, Z.; Wang, H.; Qin, W.; Wang, W.; Wang, Y.; Wang, Z.; Ebrahimi, S.; Wang, H. (2024): Continual Learning of Large Language Models: A Comprehensive Survey (No. arXiv:2404.16789). DOI: 10.48550/arXiv.2404.16789

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. (2024): AI models collapse when trained on recursively generated data. In: Nature 631(8022), S. 755–759. DOI: 10.1038/s41586–024–07566-y

Sun, J.; Zheng, C.; Xie, E.; Liu, Z.; Chu, R.; Qiu, J.; Xu, J.; Ding, M.; Li, H.; Geng, M.; Wu, Y.; Wang, W.; Chen, J.; Yin, Z.; Ren, X.; Fu, J.; He, J.; Yuan, W.; Liu, Q.; Liu, X.; Li, Y.; Dong, H.; Cheng, Y.; Zhang, M.; Heng, P. A.; Dai, J.; Luo, P.; Wang, J.; Wen, J.-R.; Qiu, X.; Guo, Y.; Xiong, H.; Liu, Q. Li, Z. (2023, December 17). A Survey of Reasoning with Foundation Models. DOI: 10.48550/arXiv.2312.11562

Tuschling, A.; Sudmann, A.; Dotzler, B. J. (Hg.) (2023): ChatGPT und andere »Quatschmaschinen«: Gespräche mit Künstlicher Intelligenz. Bielefeld. DOI: 10.14361/9783839469088

Vafa, K.; Rambachan, A.; Mullainathan, S. (2024): Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function (No. arXiv:2406.01382). DOI: 10.48550/arXiv.2406.01382

Ven, G. M. van de; Soures, N.; Kudithipudi, D. (2024): Continual Learning and Catastrophic Forgetting (No. arXiv:2403.05175; Version 1). DOI: 10.48550/arXiv.2403.05175

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. (2023): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (No. arXiv:2201.11903). DOI: 10.48550/arXiv.2201.11903

Wu, X.; Zhao, H.; Zhu, Y.; Shi, Y.; Yang, F.; Liu, T.; Zhai, X.; Yao, W.; Li, J.; Du, M.; Liu, N. (2024): Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era (No. arXiv:2403.08946). DOI: 10.48550/arXiv.2403.08946

Xu, L.; Su, Z.; Yu, M.; Xu, J.; Choi, J. D.; Zhou, J.; Liu, F. (2024): Identifying Factual Inconsistencies in Summaries: Grounding Model Inference via Task Taxonomy (No. arXiv:2402.12821). DOI: 10.48550/arXiv.2402.12821

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. (2024): A Survey on Multimodal Large Language Models. National Science Review 11(12). DOI: 10.1093/nsr/nwae403