Grosse Sprachmodelle

Siegfried Handschuh

Der Artikel gibt einen umfassenden Überblick über den aktuellen Stand der Forschung zur generativen KI und insbesondere grossen Sprachmodellen (Large Language Models, LLMs). Es werden die Architektur, das Training und die emergenten Fähigkeiten von LLMs wie GPT-3 erläutert. Grosse Sprachmodelle basieren auf neuronalen Netzen und werden auf riesigen Textdatenmengen trainiert. Dabei lernen sie, basierend auf dem bisherigen Textverlauf das jeweils nächste Wort vorherzusagen. Obwohl dies eine einfache Aufgabe ist, ermöglicht dies komplexe sprachliche Fähigkeiten. Mit zunehmender Modellgrösse zeigen LLMs dabei unerwartete emergente Fähigkeiten wie Textzusammenfassung, mathematische Operationen oder räumliches Denken.
Allerdings haben LLMs auch Schwächen wie die Tendenz zum Fabulieren bei Wissenslücken und mangelnde Kohärenz. Aktuell gibt es rasante Fortschritte durch neue Modelle wie GPT-3 und ChatGPT. Zukünftige Entwicklungen müssen ethische Aspekte berücksichtigen. Insgesamt eröffnen grosse Sprachmodelle faszinierende Möglichkeiten, aber weitere Forschung ist nötig. Der Artikel liefert eine umfassende Übersicht zu Chancen und Herausforderungen dieses rasanten Technologiefeldes.

L'article donne un aperçu complet de l'état actuel de la recherche sur l'IA générative, en particulier sur les grands modèles de langage (Large Language Models, LLMs). Il explique l'architecture, l'apprentissage et les capacités émergentes des LLM comme GPT-3. Les grands modèles linguistiques sont basés sur des réseaux neuronaux et sont entraînés sur d'énormes quantités de données textuelles. Ils apprennent ainsi à prédire le mot suivant en se basant sur le déroulement du texte en amont. Il s'agit d'une tâche simple, mais elle permet d'acquérir des compétences linguistiques complexes. Avec l'augmentation de la taille du modèle, les LLM montrent des capacités émergentes inattendues telles que les résumés de textes, les opérations mathématiques ou le raisonnement spatial.
Toutefois, les LLM présentent aussi des faiblesses, comme la tendance à fabuler en cas de lacunes dans les connaissances et le manque de cohérence. Actuellement, les progrès sont rapides grâce à de nouveaux modèles comme GPT-3 et ChatGPT.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 8 (2024) — 10.18755/iw.2024.3

Les développements futurs devront tenir compte des aspects éthiques. Dans l'ensemble, les grands modèles linguistiques ouvrent des possibilités fascinantes, mais des recherches supplémentaires sont nécessaires. Cet article fournit un aperçu complet des opportunités et des défis de ce domaine technologique en plein essor.

The article provides a comprehensive overview of the current state of research on generative AI and in particular large language models (LLMs). It explains the architecture, training and emergent capabilities of LLMs such as GPT-3. Large language models are based on neural networks and are trained on huge amounts of text data. In doing so, they learn to predict the next word based on the previous text. Although this is a simple task, it enables complex linguistic abilities. With increasing model size, LLMs show unexpected emergent abilities such as text summarisation, mathematical operations or spatial reasoning.
However, LLMs also have weaknesses such as a tendency to fabricate when there are knowledge gaps or a lack of coherence. Rapid progress is currently being made with new models such as GPT-3 and ChatGPT. Future developments must take ethical aspects into account. Overall, large language models open up fascinating possibilities, but more research is needed. This article provides a comprehensive overview of the opportunities and challenges in this fast-paced field of technology.

1 Einleitung

Grosse Sprachmodelle (engl.: Large Language Models, LLMs) sind derzeit die Flaggschiffe der generativen künstlichen Intelligenz, und insbesondere mit dem Erscheinen von ChatGPT am 30. November 2022[1] hat sich eine soziotechnische Zeitenwende eingeleitet. Eine Zeitenwende, wie wir sie nur alle 30–40 Jahre erleben und wie wir sie zuletzt mit der Einführung des World Wide Web gesehen haben. Es ist absehbar, dass generative KI im Allgemeinen und Sprachmodelle im Besonderen die Art und Weise verändern, wie wir forschen, lernen, kommunizieren und wie wir arbeiten, kurz, wie wir leben.

In unserer Forschung beschäftigen wir uns seit über 20 Jahren mit der Semantik von Wörtern. Anfangs verwendeten wir formale Systeme, d. h. grammatikalische Parser, Ontologien und wörterbuchähnliche Strukturen wie WordNet[2]. Dieser Ansatz funktionierte, wurde jedoch immer umfangreicher, je kom-

1 GPT-3.5. 15.03.2022 https://platform.openai.com/docs/model-index-for-researchers
2 WordNet: A lexical database for English. Communications of the ACM, 38(11), 39–41. https://doi.org/10.1145/219717.219748

plexer die Sätze waren. Andererseits versagten diese Parser, wenn sich die Sprachen nicht an grammatikalische Regeln hielten oder Neologismen verwendeten, Phänomene, die wir oft in sozialen Netzwerken erleben.

In den vergangenen 14 Jahren hat sich in unserer Forschung und dann auch innerhalb der wissenschaftlichen Gemeinschaft eine verstärkte Hinwendung zu statistischen Modellen vollzogen, insbesondere unter Einsatz der sogenannten Verteilungssemantik (engl.: Distributional Semantics)[3]. Diese Methodik entwickelt und untersucht die Charakteristika sprachlicher Elemente basierend auf ihren Verteilungseigenschaften innerhalb umfangreicher Sprachdatensätze. Es hat sich gezeigt, dass sich mit dieser Methodik die Phänomene der Semantik deutlich präziser beschreiben lassen. In dieser Analogie können Wörter metaphorisch als Moleküle betrachtet werden, die mit einer bestimmten Wahrscheinlichkeit interagieren. Ähnlich verhält es sich mit Metaphern und Idiomen, die als Muster verstanden werden können, die sich im Laufe der Zeit im Sprachgebrauch herausbilden.

In der Machine-Learning-Community hat sich mit den Word Embeddings[4]ein mit der Verteilungssemantik stark verwandter Ansatz durchgesetzt, der wiederum die Basis für weitere Entwicklungen im Bereich des Sprachverstehens bildete. Unter-Word Embeddings versteht man ein numerischer Zahlenvektor, der die Semantik eines Wortes in seinem gegebenen Kontext beschreib.

Eine entscheidende Entwicklung hier war das Transformer-Modell von Google, bekannt durch das Paper „Attention Is All You Need“ [5], das Word-Embeddings mit Attention-Mechanismen kombiniert. Während die Word-Embeddings speziell das Semantikproblem der Sprache adressieren, lösen die Attention-Mechanismen weitere Schlüsselprobleme der Computerlinguistik, wie beispielsweise Koreferenzauflösung, Wortsinn-Desambiguierung oder das Verarbeiten elliptischer Konstruktionen.

Generative Pre-trained Transformer 1 (GPT-1) war 2018 das erste der grossen Sprachmodelle von OpenAI[6], nachdem Google ein Jahr zuvor die Transformer-Architektur erfunden hatte. Im Gegensatz zum Transformer-Modell in

3 Baroni, Marco; Lenci, Alessandro: Distributional Memory: A General Framework for Corpus-Based Semantics. Computational Linguistics, 2010. Band 36(4), S. 673–721. https://doi.org/10.1162/coli_a_00016

4 Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeffrey: Distributed Representations of Words and Phrases and Their Compositionality. In: Advances in Neural Information Processing Systems, 2013. Band 26, S. 3111–3119.

5 Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia: Attention Is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, 2017, S. 6000-6010.

seiner ursprünglichen Form, das sowohl Encoder als auch Decoder umfasst, konzentriert sich GPT auf die Generierung von Text und verwendet daher nur den Decoder-Teil der Transformer-Architektur. GPT-2 wurde 2019[7]veröffentlicht und GPT-3 folgte 2020[8]. Während GPT-1 und GPT-2 bereits in der wissenschaftlichen Community Beachtung fanden, erzielte GPT-3 aufgrund seiner emergenten Fähigkeiten deutlich mehr Aufmerksamkeit. Insbesondere durch die Arbeiten von OpenAI beim Feintuning von GPT-3 in Form von Instruct-GPT[9], bei welchem dem System beigebracht wurde, menschlich formulierte Anweisungen zu befolgen, wurde ein wichtiger Meilenstein erreicht auf dem Weg zu ChatGPT.

In den letzten Jahren haben grosse Sprachmodelle für Aufsehen in der KI-Forschung gesorgt. Diese Modelle zeigen beeindruckende Fähigkeiten in der Verarbeitung natürlicher Sprache und stellen einen wichtigen Fortschritt auf dem Gebiet der künstlichen Intelligenz dar[10]. Allerdings weisen LLMs auch bestimmte Einschränkungen und Risiken auf, die in der wissenschaftlichen Diskussion kontrovers betrachtet werden. In diesem Beitrag soll ein Überblick über den aktuellen Forschungsstand zu LLMs gegeben und diskutiert werden, welche Herausforderungen bestehen und wie die weitere Entwicklung aussehen könnte.

6 Radford, Alec; Wu, Jeffrey; Child, Rewon; Luan, David; Amodei, Dario; Sutskever, Ilya: Language Models are Unsupervised Multitask Learners. OpenAI Technical Report, San Francisco, CA. 2019. Verfügbar unter: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (abgerufen am 02.11.2023

7 Radford et al., 2019

8 Brown, Tom; Mann, Ben; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared D.; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Greg; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, David; Wu, Jeff; Winter, Clemens; et al.: Language Models are Few-Shot Learners. In: Larochelle, Hugo; Ranzato, Marc'Aurelio; Hadsell, Raia; Balcan, Maria-Florina; Lin, Hsuan-Tien (Hrsg.): Advances in Neural Information Processing Systems 33 (NIPS 2020). Curran Associates, Inc., 2020, S. 1877–1901.

9 Ouyang, Long; Wu, Jeff; Jiang, Jennie; Almeida, Daniel; Wainwright, Collin L.; Mishkin, Pamela; Zhang, Chen; Agarwal, Sandhini; Slama, Kamil; Ray, Alex; Schulman, John et al.: Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155, 2022.

10 Bommasani, Rishi; Hudson, Daniel A.; Adeli, Ehsan; Altman, Russ; Arora, Simran; von Arx, Sam; ...; Bohg, Jeannette: On the Opportunities and Risks of Foundation Models. arXiv Preprint arXiv:2108.07258, letzte Überarbeitung 12. Juli 2022

2 Architektur grosser Sprachmodelle

In den vergangenen zwei Dekaden hat sich die linguistische Forschung zunehmend auf statistische Methoden der Semantikanalyse fokussiert[11]. Im Gegensatz zu früheren Ansätzen, welche Sprache als regelbasiertes System modellierten, hat sich gezeigt, dass sich linguistische Phänomene effektiver mit probabilistischen Modellen abbilden lassen, bei denen Wörter (streng genommen sind es Subwörter bzw. Subtokens, aber wir sprechen hier der Einfachheit halber von Wörtern) als Einheiten betrachtet werden, die sich mit einer bestimmten Wahrscheinlichkeit verbinden. Über die Analyse von Co-Occurrence-Mustern, also dem gemeinsamen Auftreten von Wörtern in sprachlichen Kontexten, können semantische Strukturen modelliert werden.

Diese Co-Occurrence-Modelle nennen wir Distributional-Semantics[12], und die wissenschaftliche Community hat begonnen, im grossen Stil mit solchen Vektorräumen zu arbeiten. Wenn ein Wortschatz beispielsweise 25.000 Wörter umfasst, hat dieser Vektorraum 25.000 Dimensionen. Wir Menschen sind auf drei Dimensionen beschränkt, weshalb wir uns 25.000 Dimensionen kaum vorstellen können. Oft werden auch komprimierte Vektorräume mit reduzierten Dimensionen erstellt; eine solche Form sind Word Embeddings[13], die von 300 (Word2Vec), über 1024 (BERT-Large) bis 12,288 (GPT-3) Dimensionen reichen können.

11 Baroni/Lenci, 2010
12 Baroni/Lenci, 2010
13 Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeffrey: Distributed Representations of Words and Phrases and Their Compositionality. In: Advances in Neural Information Processing Systems, Band 26, 2013, S. 3111-3119. und auch Pennington, Jeffrey; Socher, Richard; Manning, Christopher D.: GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, S. 1532–1543.

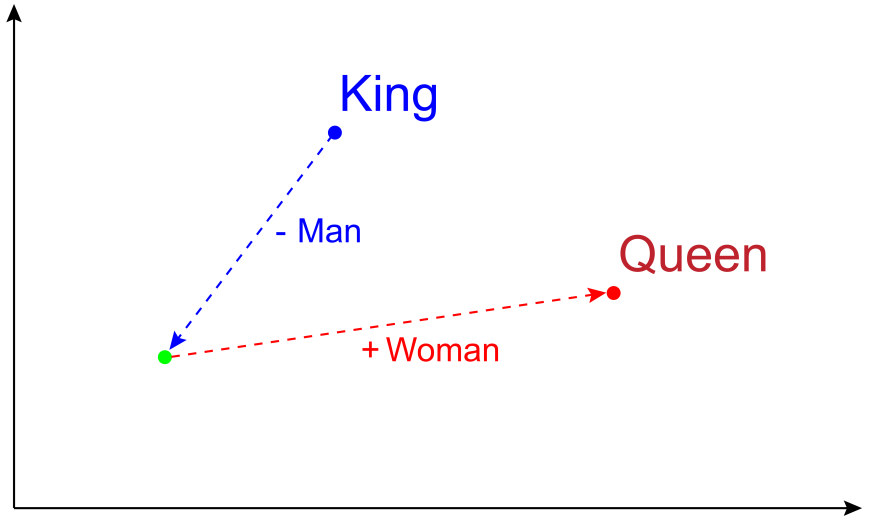

Abb. 1. Ähnlichkeiten zwischen Wörtern im Vektorraum

Es zeigt sich, dass dieser Vektorraum Ähnlichkeiten zwischen Wörtern gut darstellen kann, zum Beispiel dass "König" und "Königin" ähnliche Wörter (vgl. Abbildung 1) sind. Ein solcher Vektorraum kann auch Assoziationen gut abbilden, zum Beispiel dass "König" mit "Schloss" und "Krönung" verbunden ist und dass "Charles" ein König ist (siehe Abbildung 2). Mit diesem Modell können bestehende sprachliche Phänomene hervorragend dargestellt werden. Dies ermöglicht eine Verallgemeinerung unseres Wissens über Sprache. Dabei werden verschiedene Arten von sprachlichen Phänomenen generalisiert, einschliesslich des Weltwissens, des grammatikalischen Wissens (siehe Abbildung 3) und des Wissens über Metaphern. Es ist jedoch zu beachten, dass eine solche Generalisierung nur dann stattfindet, wenn diese Metaphern in den Daten in ausreichender Menge vorhanden sind.

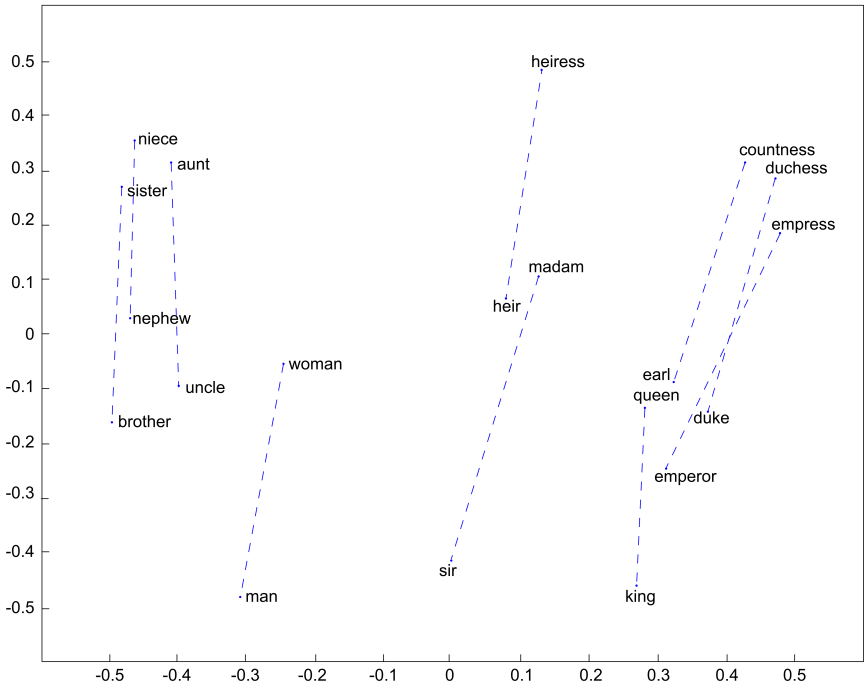

Abb. 2. Wortassoziationen und Wortähnlichkeiten im Vektorraummodell

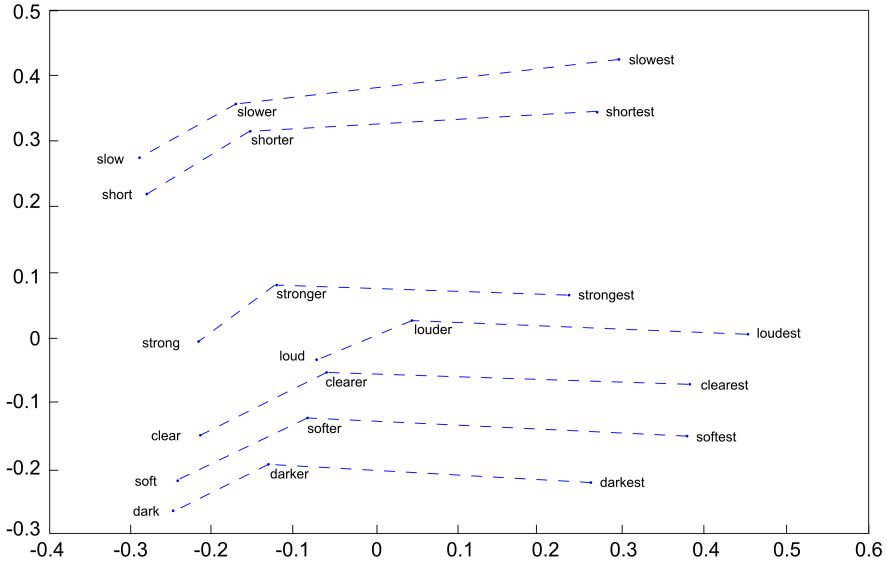

Abb. 3. Abbildung grammatikalischer Strukturen, hier Superlative, im Vektorraummodell

Aktuelle Forschung im Bereich der Computerlinguistik setzt vermehrt vektorbasierte Repräsentationen von Subwörtern ein, bei denen jedes Wort durch einen hochdimensionalen Vektor repräsentiert wird, der seine Bedeutung codiert[14]. Die Dimensionalität dieser Vektorräume ist dabei sehr hoch, da sie der Anzahl der verschiedenen Wörter entspricht. Jedoch hat sich gezeigt, dass sich die Komplexität durch Reduktion auf einige Hundert Dimensionen reduzieren lässt, ohne signifikanten Informationsverlust bezüglich linguistischer Phänomene. Neuronale Sprachmodelle wie ELMo[15], BERT[16]oder GPT-3[17] nutzen derartige vektorbasierte Wortrepräsentationen als Grundlage.

14 Mikolov et al., 2013.

15 Peters, Matthew E.; Neumann, Mark; Iyyer, Mohit; Gardner, Matt; Clark, Christopher; Lee, Kenton; Zettlemoyer, Luke: Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, S. 2227-2237.

16 Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019

Insbesondere für die Generalisierung über sprachliches Wissen ist eine grosse Menge an Trainingsdaten erforderlich. So wurden für das Training von GPT-3 etwa 500 Milliarden Wörter (eigentlich Tokens) aus diversen Quellen wie Websites, Literatur, Wikipedia oder Fachtexten verwendet[18] – eine Datenmenge, die schätzungsweise 5000 Mal größer ist als die Textmenge, die ein Mensch im Laufe seines Lebens lesen kann.

Die Rohtexte werden dabei mittels neuronaler Netze in einen hochdimensionalen Vektorraum (mit an die 1000 Zahlenwerten pro Wort) transformiert, so dass die Textsammlung durch ca. 500 Billionen Zahlen repräsentiert wird. Die Anzahl der Parameter bei GPT-3 beträgt hingegen 175 Milliarden. Man kann man also feststellen, dass die Anzahl der Parameter etwa ein Dreitausendstel des Datenvolumens beträgt, d. h., die Speicherfähigkeit des Netzes liegt deutlich unter der Trainingsmenge. Daher muss über die gelesenen Texte generalisiert werden. Auch wird die Quellenangabe der Rohtexte nicht im Netzwerk gespeichert und ist daher auch nicht mehr rekonstruierbar.

Nun erfolgt das Training auf einer GPU. Eine GPU ist ein Prozessor, der ursprünglich für Computerspiele entwickelt wurde, da er in der Lage ist, schnelle lineare Algebra-Berechnungen durchzuführen, die für Computergrafik benötigt werden. Es hat sich herausgestellt, dass sich diese Art von Berechnungen auch gut für die Vektorraummodelle der Sprachmodelle eignet.

Würde man das Training für GPT-3 mit einer Standard-GPU erbringen wollen, einer NVIDIA Tesla V100 GPU, würde es nach einer Schätzung von LambdaLabs[19] 355 Jahre dauern. Dies veranschaulicht die enorme Rechenkapazität, die für solche Modelle erforderlich ist. Selbst mit zwei NVIDIA DGX-2 (je 16 Tesla V100 GPUs), die uns an der Universität St.Gallen zur Verfügung stehen, würden wir geschätzt 11 Jahre benötigen. Bei OpenAI soll die Berechnung auf 520 GPUs gelaufen sein und wurde in etwa 250 Tagen abgeschlossen. Die Schätzungen für die Trainingskosten von GPT-3 liegen zwischen 1,8 Millionen[20]und

17 Brown, Tom; Mann, Ben; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared D.; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Greg; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, David; Wu, Jeff; Winter, Clemens; ...; Amodei, Dario: Language Models are Few-Shot Learners (GPT-3). In: Larochelle, Hugo; Ranzato, Marc'Aurelio; Hadsell, Raia; Balcan, Maria-Florina; Lin, Hsuan-Tien (Hrsg.): Advances in Neural Information Processing Systems, Band 33, 2020, S. 1877–1901. Curran Associates, Inc. Verfügbar unter: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

18 Brown et al., 2020

19 https://lambdalabs.com/blog/demystifying-gpt-3

20 Stanford University, Human-Centered AI Institute: Annual Report, 2023. Verfügbar unter: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (abgerufen am 02.11.2023)

4,6 Millionen US-Dollar[21]. Nicht nur für die Berechnung, sondern auch um GPT auszuführen, ist mindestens eine Grafikkarte erforderlich. Für den Einsatz von GPT-3.5 – also für die Bearbeitung von Anfragen, nicht für das Training – sollen 600 GB an Speicher sowie eine dedizierte GPU mit 250 GB VRAM erforderlich sein. Daher ist sowohl die Entwicklung eines Sprachmodells wie GPT als auch dessen permanente Nutzung mit finanziellen Aufwendungen verbunden.

Laut einem Bericht der Nachrichtenwebsite Semafor[22] verfügt GPT-4, das Nachfolgemodell von GPT-3, welches am 14. März 2023 vorgestellt wurde, über eine Billion Parameter - etwa sechs Mal mehr als sein Vorgänger GPT-3. Andere Quellen gehen jedoch davon aus, dass GPT-4 aus acht Modellen mit jeweils 220 Milliarden Parametern besteht, was zusammengenommen rund 1,76 Billionen Parameter ergeben würde. Sam Altman, CEO von OpenAI, dem Unternehmen hinter GPT-4, sagte, dass die Kosten für das Training von GPT-4 mehr als 100 Millionen Dollar [23] betragen sollen. Die genaue Anzahl der Parameter von GPT-4 hat OpenAI bislang nicht offiziell bestätigt.

Zusammenfassend lässt sich sagen, dass das Training der Modelle unter hohem Rechenaufwand auf leistungsfähiger GPU-Hardware erfolgt. Für die beiden GPT-Modelle wurden über einen Zeitraum von mehreren Monaten hinweg Hunderte von Grafikprozessoren parallel eingesetzt, was Kosten in Millionenhöhe verursachte.

3 Die Vorhersage des nächsten Wortes

Große Sprachmodelle zeigen die bemerkenswerte Fähigkeit, komplexe Fragen zu beantworten, obwohl sie ursprünglich nicht für diese Aufgabe entwickelt wurden. Vielmehr handelt es sich um generative KI-Systeme, deren Ziel die Produktion von fortlaufendem Text auf Basis gegebener Texteingaben, sogenannter Prompts[24], ist.

21 https://lambdalabs.com/blog/demystifying-gpt-3

22 Albergotti, Reed: The Secret History of Elon Musk, Sam Altman, and OpenAI. In: Semafor, 2023. Verfügbar unter: https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai (abgerufen am 02.11.2023)

23 Knight, Will: OpenAI's CEO Says the Age of Giant AI Models Is Already Over. In: Wired, 2023. Verfügbar unter: https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/ (abgerufen am 02.11.2023)

24 Reynolds, Laura; McDonell, Kevin: Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In: CHI EA '21: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Artikel Nr. 314, 2021, S. 1–7. Verfügbar unter: https://doi.org/10.1145/3411763.3451760.

Der grundlegende Algorithmus dieser Systeme ist die sequenzielle Vorhersage des jeweils nächsten Wortes, basierend auf den vorhergehenden Wörtern und dem Kontext[25]. Während des Trainings werden die Modellparameter so optimiert, dass die Vorhersagewahrscheinlichkeit korrekter Fortsetzungen maximiert wird. Beispielsweise könnte nach dem Prompt «Fieber wird mit einem ____ gemessen» mit hoher Wahrscheinlichkeit das Wort «Thermometer» vorhergesagt werden.

Bei ambiguen Fällen, wie «Ich fahre mit dem ____ ins Büro», gibt es mehrere mögliche Fortsetzungen wie «Auto» oder «Fahrrad» mit gewissen Wahrscheinlichkeiten. Für den Satzbeginn «Bern ist die _____» wäre «Hauptstadt» eine plausible Ergänzung mit beispielsweise 16 % Vorhersagewahrscheinlichkeit (nach GPT-2), während auch Alternativen wie «beste», «erste» oder «einzige» möglich sind, wenn auch mit geringerer Wahrscheinlichkeit.

Obwohl die Vorhersage des nächsten Wortes eine einfache Aufgabe darstellt, ermöglicht dies in der Summe, komplexe Phänomene der Sprachproduktion zu modellieren. Die Fähigkeit der Fragebeantwortung emergiert dabei aus dem generationsbasierten Ansatz, da plausibel erscheinende Fortsetzungen eines Dialogs generiert werden. Die intrinsische Komplexität des zugrunde liegenden linguistischen Wissens ergibt sich aus der schieren Menge der Trainingsdaten.

4 Emergente Fähigkeiten

Aktuelle Forschungsergebnisse zeigen, dass grossskalige Sprachmodelle mit zunehmender Modellgrösse emergente Fähigkeiten[26] entwickeln, mit denen die ForscherInnen ursprünglich nicht gerechnet hatten. Manche dieser Fähigkeiten skalieren linear mit der Größe des Modells und der Menge der Trainingsdaten, während andere plötzlich und unerwartet auftreten.

Einige dieser emergenten Fähigkeiten, wie zum Beispiel Textzusammenfassungen zu generieren, wurden nach ihrer Entdeckung gezielt verstärkt und trainiert[27]. Des Weiteren konnte beobachtet werden, dass mathematische Fähigkeiten mit der Modellgrösse linear ansteigen. Dabei führen Sprachmodelle keine tatsächlichen Berechnungen durch, sondern approximieren mathemati-

25 Vaswani et al., 2017.

26 Wei, Jason; Tay, Yi; Bommasani, Rishi; Raffel, Colin; Zoph, Barret; Borgeaud, Sebastian; Yogatama, Dani; Bosma, Maarten; Zhou, David; Metzler, Donald; Chi, Ed H.; et al.: Emergent Abilities of Large Language Models. In: Transactions on Machine Learning Research, 8/2022.

sche Operationen basierend auf zuvor gesehenen Beispielen. Generell generieren grosse Sprachmodelle Antworten auf Fragen nicht durch echtes Verständnis, sondern interpolieren neue Texte aus vorherigen Antworten auf ähnliche Fragen.

Abb. 4. Modellgrösse und emergente Fähigkeiten der Sprachmodelle

Besonders hervorzuheben sind plötzlich auftretende Fähigkeiten, sogenannte *discontinuous improvements*, wie sie unter anderem beim Google PaLM Model[28] beobachtet wurden. Ab einer bestimmten Modellgröße können Sprachmodelle beispielsweise Abläufe und Prozesse kausal korrekt ordnen, wie etwa die Schritte des Trinkens aus einer Flasche.

Abbildung 4 zeigt die Relation zwischen Modellgrösse und zunehmenden emergenten Fähigkeiten. So zeigen große Sprachmodelle trotz fehlender sensorischer Wahrnehmung die emergente Fähigkeit zum räumlichen Denken. Nachdem das Modell ausreichend Beschreibungen von Räumen gelesen hat, kann es in virtuellen Umgebungen navigieren und sogar Brettspiele spielen. Dies widerlegt die bisherige Annahme in der KI-Forschung, dass ein Körper für den Er-

27 Taylor, Richard; Kardas, Mikołaj; Cucurull, Guillem; Scialom, Thomas; Hartshorn, Alex; Saravia, Erick; Poulton, Alex; Kerkez, Vladan; Stojnic, Radoslav: Galactica: A Large Language Model for Science. arXiv:2211.09085, 2022. Verfügbar unter: http://arxiv.org/abs/2211.09085; Ouyang, Long; Wu, Jeff; Jiang, Jennie; Almeida, Daniel; Wainwright, Collin L.; Mishkin, Pamela; Zhang, Chen; Agarwal, Sandhini; Slama, Kamil; Ray, Alex; Schulman, John; et al.: Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155, 2022.

28 Chowdhery, Aakanksha; Narang, Sharan; Devlin, Jacob; Bosma, Maarten; Mishra, Gaurav; Roberts, Adam; Barham, Paul; Chung, Hyung Won; Sutton, Charles; Gehrmann, Sebastian; Schuh, Parker; Shi, Katherine; Tsvyashchenko, Sasha; Maynez, Joshua; Rao, Abhik; Barnes, Parker; Tay, Yi; Shazeer, Noam; Prabhakaran, Vinod; ...; Fiedel, Norman: PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311, 2022. Verfügbar unter: http://arxiv.org/abs/2204.02311.

werb von Weltwissen notwendig sei. Vielmehr scheinen sprachliche Beschreibungen auszureichen.

Zusammengefasst legen diese Ergebnisse nahe, dass grossskalige Sprachmodelle weitaus mehr Fähigkeiten entwickeln als ursprünglich angenommen. Sowohl lineares Wachstum bekannter als auch plötzliches Auftreten neuer Fähigkeiten konnten beobachtet werden. Dies wirft interessante Fragen für die weitere Erforschung emergenter Phänomene in der KI auf.

5 Prompt Engineering

Wie gerade erwähnt weisen grosse Sprachmodelle eine Vielzahl von Fähigkeiten auf, die durch entsprechende Texteingaben („Prompts“) aktiviert werden können. Dies lässt sich dadurch erklären, dass diese Modelle auf Wahrscheinlichkeiten basieren und darauf trainiert wurden, Anweisungen zu befolgen und Muster zu vervollständigen. Effektive Prompts beginnen idealerweise mit einer klaren Anweisung wie „Fasse zusammen“ oder „Erzeuge“ und grenzen den Möglichkeitsraum durch Kontextinformationen ein[29]. Abschließend hilft ein Marker, die gewünschte Antwort direkt auszulösen. Mittlerweile existieren umfangreiche Prompt-Bibliotheken als Hilfestellung[30].

Einige ExpertInnen spekulieren, dass Prompt Engineering klassisches Programmieren ersetzen wird, jedoch herrscht hierzu noch Uneinigkeit. Zweifelsfrei können Sprachmodelle aber die Programmierung unterstützen, insbesondere wenn implizites Wissen abgerufen werden muss[31].

Aktuelle Studien untersuchen auch den pädagogischen Einsatz von Sprachmodellen. Laut Seufert et al.[32] können zwei effektive Strategien sein: 1) Generierung vieler Beispiele zur Veranschaulichung komplexer Konzepte und 2) schrittweise Erklärungen unter Berücksichtigung des Vorwissens. Durch eine klare, präzise Schreibweise kann das Modell als Tutor agieren. Abstrakte Konzepte lassen sich durch Analogien und detaillierte Definitionen verständlich vermitteln.

29 Reynolds/McDonell, 2021.
30 https://quickref.me/chatgpt
31 Chen, Mingyuan; Tworek, Jakub; Jun, He; Qi, Qinyuan; de Nobrega, Raphael Vasconcelos; Jain, Supriya; ...; Kiela, Douwe: Evaluating Large Language Models Trained on Code. arXiv Preprint arXiv:2107.03374, 2021
32 Seufert, Simon; Eberle, Franziska; Handschuh, Sebastian: Orientierung und erste Empfehlungen für das Gymnasium, 2023. Verfügbar unter: https://www.alexandria.unisg.ch/handle/20.500.14171/118501.

Zusammengefasst legen diese Befunde nahe, dass grossskalige Sprachmodelle durch umfangreiches Beispielmaterial und strukturierte Präsentation einen wertvollen Beitrag u. a. zum effektiven Lernen leisten können. Weitere Untersuchungen sind jedoch nötig, um die konkreten Möglichkeiten und Grenzen zu bestimmen.

6 Schwächen und Herausforderungen

Trotz ihrer beeindruckenden Fähigkeiten weisen die Sprachmodelle auch bestimmte systemimmanente Schwächen auf. Laut Bommasani et al.[33]generalisieren diese Modelle lediglich auf Basis der Trainingsdaten, und ihre Antworten sind darauf ausgerichtet, möglichst plausibel zu wirken, unabhängig von der tatsächlichen Korrektheit. Prinzipiell liefern sie zu jeder Frage eine Antwort, sofern sie nicht durch Regelwerke eingeschränkt werden .

Das Antwortverhalten dieser Modelle ähnelt in vielen Aspekten eher menschlicher Intuition als einer logisch-analytischen Herangehensweise, wie sie lange Zeit als Idealbild künstlicher Intelligenz galt. Stattdessen tendieren Sprachmodelle dazu, bei Wissenslücken spekulative Antworten zu generieren, anstatt Unsicherheit zuzugeben – ein als „Halluzination“ bezeichnetes Phänomen[34]. Der Effekt wird auch als „Fabulieren“ bezeichnet.

Im Gegensatz zu einer Datenbank können Sprachmodelle die Trainingsdaten nicht einzeln abrufen, sondern bilden Verallgemeinerungen über die Trainingsdaten. Dies kann dazu führen, dass die generierten Antworten eher durchschnittliche Eigenschaften oder Interpolationen widerspiegeln als tatsächliche Fakten. Infolgedessen kann das Modell in konkreten Anwendungsfällen falsche Antworten produzieren, auch wenn die Antwort insgesamt kohärent und plausibel erscheint. Die Neigung zum Fabulieren ist demnach eine inhärente Limitation von Sprachmodellen, die auf der Notwendigkeit beruht, komplexe Information zu generalisieren und dadurch Toleranz gegenüber verrauschten oder spärlichen Daten zu erlangen. Weitere Forschung ist nötig, um das Fabulieren zu reduzieren und die Fähigkeit der Modelle zu verfeinern, die Genauigkeit

33 Bommasani, Rishi; Hudson, David A.; Adeli, Ehsan; Altman, Robert; Arora, Sumeet; von Arx, Sina; ...; Bohg, Jeannette: On the Opportunities and Risks of Foundation Models, 2022. Letzte Überarbeitung am 12. Jul 2022. arXiv Preprint arXiv:2108.07258.

34 Ji, Zhilin; Lee, Namhoon; Frieske, Robert; Yu, Tao; Su, Dawei; Xu, Yiren; Ishii, Etsuko; Bang, Youngjoon J.; Madotto, Andrea; Fung, Pascale: Survey of Hallucination in Natural Language Generation. In: ACM Computing Surveys, 55(12), 2022, S. 248 (DOI: 10.1145/3571730).

von Aussagen selbst einzuschätzen und gegebenenfalls Unsicherheit zu signalisieren.

Grossskalige Sprachmodelle, die auf Methoden wie Next-Word-Prediction basieren, weisen mitunter Mängel in der strukturellen Kohärenz und Konsistenz generierter Texte auf[35]. Empirische Analysen zeigen, dass die von Large Language Models (LLMs) produzierten Antworten gelegentlich redundant, widersprüchlich oder inkohärent sind. LLMs neigen dazu, ausführlichere Antworten zu generieren als nötig, mit wiederholenden Textpassagen und wissenschaftlichen Aussagen, die nicht logisch miteinander verknüpft sind. Sogar direkte Widersprüche in ein und derselben Antwort treten auf. Diese Mängel sind eine direkte Folge des Next-Word-Prediction-Ansatzes, welcher Wörter sequenziell und ohne übergeordnete Textplanung vorhersagt. Dadurch fehlt den generierten Antworten eine kohärente semantische und argumentative Struktur. Um dieses Defizit zu beheben, sind Erweiterungen der bestehenden Modelle um Komponenten für globale Textplanung und -kohärenz vielversprechende Ansatzpunkte für die weitere Forschung.

Grossskalige Sprachmodelle wie LLMs generieren Text, ohne die Herkunft der enthaltenen Informationen anzugeben. Dies stellt ein bekanntes Defizit ihrer systemimmanenten Funktionsweise dar[36]. Ein vielversprechender Lösungsansatz ist Retrieval-Augmented Generation (RAG). Dabei wird das Sprachmodell um eine Komponente für das Auffinden relevanter Referenzen erweitert, welche dann in die generierte Antwort integriert werden. Erste Implementierungen wie in Microsofts Suchmaschine Bing zeigen, dass RAG oft, wenn auch nicht ausnahmslos, korrekte Referenzen einfügen kann. Dadurch kann das Fabulieren, also das Generieren spekulativer Inhalte ohne faktische Grundlage, reduziert werden. Jedoch sind weitergehende Forschungsarbeiten nötig, um Referenzierungen konsistent und umfassend in grossskalige Sprachmodelle zu integrieren. Die Fähigkeit, getätigte Aussagen durch korrekte Zitate und Quellenangaben zu belegen, ist essenziell für eine vertrauenswürdige und transparente Textgenerierung durch KI.

35 McCoy, Ryan T.; Yao, Shiyue; Friedman, David; Hardy, Marcus; Griffiths, Thomas L.: Embers of Autoregression: Understanding Large Language Models through the Problem They Are Trained to Solve. Retrieved from https://arxiv.org/abs/2309.13638, 2023.

36 Lewis, Patrick; Perez, Ethan; Piktus, Aaron; Petroni, Fabio; Karpukhin, Vladimir; Goyal, Nitish; Küttler, Hannes; Lewis, Mike; Yih, Wen-tau; Rocktäschel, Tim; Riedel, Sebastian; Kiela, Douwe: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems, 33, 2020, S. 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Eine weitere Herausforderung besteht darin, dass das Training von Sprachmodellen kostspielig ist und die Ausführung der Modelle in der Regel eine Grafikkarte erfordert. Aus diesem Grund beschränkt OpenAI anfänglich den Zugang zu GPT-4 sogar für zahlende Kunden.

Zusammengefasst legen diese Erkenntnisse nahe, dass trotz aller Fortschritte weiterhin substanzieller Forschungsbedarf hinsichtlich der systematischen Schwächen grossskaliger Sprachmodelle besteht.

7 Aktuelle Entwicklungen

Die jüngsten Fortschritte bei grossskaligen Sprachmodellen (LLMs) sind von beeindruckender Dynamik geprägt. Wir durchleben gegenwärtig eine Phase rasanten Wandels, in der Innovationen, für die normalerweise mehrere Jahre benötigt werden, derzeit innerhalb weniger Monate realisiert werden.

Nachdem anzunehmen war, dass die grossen kommerziellen und proprietären Sprachmodelle (ChatGPT, Bard, Claude etc.) den Markt monopolisieren könnten, hat sich erfreulicherweise, aus Sicht der Forschung, Meta von Facebook dafür entschieden, seine Modelle öffentlich zugänglich zu machen. Dazu zählen LLaMA[37], welches im Februar 2023 veröffentlicht wurde, und später im Juli 2023 LLaMA 2[38] von Meta. LLaMA 2 verfügt über bis zu 70 Milliarden Parameter und wurde mit 2 Billionen Wörtern (Tokens) aus öffentlich verfügbaren Datenquellen trainiert.

Um Sprachmodellen Chatfähigkeit beizubringen, d. h. Instruktionen zu befolgen, gibt es neben dem aufwendigen Reinforcement Learning[39] mit hohem menschlichem Aufwand (eingesetzt bei ChatGPT und LLaMA 2) auch folgende Möglichkeit: große Modelle, einschließlich kommerzieller Modelle, als Lehrer einzusetzen, um kleineren Schülermodellen ähnliche Fähigkeiten zu vermitteln[40]. Dieser Ansatz, der von Stanford beim Erstellen des Alpaca-Modells an-

37 Touvron, Hervé; Lavril, Thibaut; Izacard, Gabriel; Martinet, Xavier; Lachaux, Marie-Anne; Lacroix, Timothée; Lample, Guillaume: LLaMA: Open and Efficient Foundation Language Models, 2023. Abgerufen von https://arxiv.org/abs/2302.13971.

38 Touvron et al., 2023

39 Ziegler, David M.; Stiennon, Nicolas; Wu, Jeff; Brown, Tom B.; Radford, Alec; Amodei, Dario; Christiano, Paul; Irving, Geoffrey: Fine-Tuning Language Models from Human Preferences. arXiv Preprint arXiv:1909.08593, 2019.

40 Wang, Yang; Kordi, Yasaman; Mishra, Suvam; Liu, Alan; Smith, Noah A.; Khashabi, Daniel; Hajishirzi, Hannaneh: Self-Instruct: Aligning Language Models with Self-Generated Instructions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Hrsg. von A. Rogers, J. Boyd-Graber, & N. Okazaki, Association for Computational Linguistics, 2023, S. 13484–13508. https://doi.org/10.18653/v1/2023.acl-long.754

gewendet wurde, kann die Trainingszeit und Kosten signifikant reduzieren[41]. Alpaca, das auf dem kleineren LLaMA-Modell basiert, wurde mithilfe von OpenAIs GPT-3 trainiert, um Instruktionen wie Zusammenfassen, Übersetzen und Strukturieren zu befolgen. Dies ermöglichte es Alpaca, Fähigkeiten ähnlich denen von ChatGPT zu erlangen, wobei das Lehrer-Schüler-Training Stanford weniger als 600 Dollar kostete.

Wie zuvor erwähnt, benötigt die Ausführung von Sprachmodellen in der Regel eine leistungsstarke, professionelle Grafikkarte und ausreichend Speicherplatz. Es gibt jedoch auch die Möglichkeit der Modellschrumpfung, genauer der Quantisierung. Dabei werden die kontinuierlichen Gewichte des neuronalen Netzes auf eine geringere Anzahl diskreter Werte abgebildet. Dies reduziert den Speicherbedarf, da die Gewichte mit weniger Bits repräsentiert werden können, und beschleunigt die Berechnungen, da mit diskreten Werten einfacher zu rechnen ist. Der Nachteil ist eine gewisse Genauigkeitseinbusse gegenüber dem Originalmodell. Durch geschickte Wahl der Quantisierungsstufen lässt sich dieser Effekt jedoch in Grenzen halten. So kann durch Quantisierung die Ausführung von Sprachmodellen mit beschränkten Ressourcen, wie auf einem Laptop, ermöglicht werden. Die Quantisierung ermöglicht also eine effizientere Ausführung von Sprachmodellen, indem die Modellgröße reduziert wird. Dies ist zwar mit Genauigkeitseinbußen verbunden, kann aber den Ressourcenbedarf deutlich senken.

Im Vergleich zur künstlichen Intelligenz arbeitet das menschliche Gehirn mit einer weit höheren Effizienz. Obwohl sich die grundlegende Architektur unserer Neuronen stark von KI-Algorithmen unterscheidet, ist das Gehirn mit einem Energieverbrauch von nur etwa 20 Watt, was rund 20 % unserer täglichen Energieaufnahme entspricht, bemerkenswert energieeffizient. Unsere Neuronen feuern Aktionspotenziale und zeigen Ermüdungserscheinungen, während Algorithmen Berechnungen kontinuierlich und ohne Ermüdung durchführen können. Mit schätzungsweise 100 Billionen Synapsen übertrifft das menschliche Gehirn aktuelle KI-Systeme noch deutlich in Komplexität und Leistungsfähigkeit. Obwohl das menschliche Gehirn noch deutlich in Komplexität und Effizienz vorne liegt, deuten neuere Studien auf mögliche Parallelen in der grundlegenden Funktionsweise hin.

41 Taori, Rohun; Gulrajani, Ishaan; Zhang, Ting; Dubois, Yann; Li, Xuezhi; Guestrin, Carlos; Liang, Percy; Hashimoto, Tsutomu B.: Stanford Alpaca: An Instruction-following LLaMA model, 2023. GitHub Repository. Abgerufen von https://github.com/tatsu-lab/stanford_alpaca.

Eine Studie von Schrimpf et al.[42] am Massachusetts Institute of Technology (MIT) untersuchte mögliche Parallelen zwischen den Algorithmusarchitekturen von Large Language Models (LLMs) und messbaren Prozessen in der menschlichen Sprachverarbeitung. Die Autoren analysierten die Aktivierungsmuster der Knoten in einem LLM, während dieses eine Wortsequenz vorhersagte. Anschliessend verglichen sie diese Aktivierungsmuster mit Messungen der Gehirnaktivität beim Menschen während sprachlicher Aufgaben. Die Ergebnisse deuten darauf hin, dass Vorhersagemechanismen, wie sie in LLMs implementiert sind, auch eine Rolle in der menschlichen Sprachverarbeitung spielen könnten. Insbesondere scheint die Fähigkeit, basierend auf dem sprachlichen Kontext das jeweils nächste Wort vorherzusagen, sowohl in LLMs als auch im menschlichen Gehirn ähnliche Prozesse anzustoßen. Die Autoren schlussfolgern, dass solche Vorhersagemechanismen eine wichtige Funktion für Sprachverständnis und -produktion haben könnten. Insgesamt liefert die Studie erste Hinweise auf mögliche neurobiologische Parallelen zwischen künstlicher und menschlicher Intelligenz im Bereich der Sprachverarbeitung.

Neben Sprache wird darüber hinaus an Foundation Models gearbeitet, die noch umfassendere Fähigkeiten versprechen. Bei Foundation Models handelt es sich um generative Modelle, die grosse Mengen multimodaler Daten integrieren und Zusammenhänge zwischen unterschiedlichen Modalitäten wie Bild, Bewegung und Sprache herstellen können. Insbesondere im Bereich der Robotik erscheinen solche Ansätze vielversprechend, um komplexe Sensomotorik und Interaktion mit der realen Welt zu erlernen. Bisher stellen für Roboter noch vergleichsweise einfache Tätigkeiten wie das Öffnen von Türen eine Herausforderung dar. Durch die Erfassung umfangreicher multimodaler Datensätze, etwa aus egozentrischen Video- und Sprachaufnahmen von Testpersonen beim Ausführen alltäglicher Tätigkeiten über längere Zeiträume, könnten jedoch detaillierte Foundation Models trainiert werden. Diese Foundation Models könnten Roboter dann befähigen, die erlernten Fähigkeiten auf neue Situationen zu übertragen und die jeweils adäquaten motorischen und sprachlichen Aktionen auszuführen. Entsprechende Fortschritte deuten darauf hin, dass die Entwicklung von Servicerobotern für den Haushaltseinsatz, die alltägliche Aufgaben robust und zuverlässig erfüllen können, in Reichweite rückt.

42 Schrimpf, Martin; Blank, Idan A.; Tuckute, Greta; Kauf, Carina; Hosseini, Eghbal A.; Kanwisher, Nancy; Tenenbaum, Joshua B.; Fedorenko, Evelina: The neural architecture of language: Integrative modeling converges on predictive processing. In: Proceedings of the National Academy of Sciences, Band 118, Nr. 45, e2105646118, 2021.

Insgesamt befinden wir uns in einer Phase rascher Weiterentwicklung auf dem Gebiet der generativen künstlichen Intelligenz, die faszinierende Anwendungsperspektiven eröffnet.

8 Zusammenfassung

Zusammenfassend lässt sich festhalten, dass grossskalige Sprachmodelle in den letzten Jahren immense Fortschritte erzielt haben und beeindruckende Fähigkeiten aufweisen. Sie stellen einen Meilenstein in der Entwicklung generativer KI dar und ermöglichen vielversprechende Anwendungen von der Wissensvermittlung bis zur Programmierunterstützung.

Allerdings gibt es auch noch substanzielle Einschränkungen, die in der wissenschaftlichen und gesellschaftlichen Diskussion kontrovers betrachtet werden. Insbesondere die Undurchsichtigkeit der Modelle, ihre Anfälligkeit für Fehler und Halluzinationen sowie potenzielle Verzerrungen aufgrund der Trainingsdaten sind kritisch zu sehen.

Aktuell durchleben wir eine Phase immensen Fortschritts, angetrieben durch innovative Modelle wie GPT-3, ChatGPT und LLaMA (Bard, Claude etc.). Durch den Zugang zu offenen Modellen und Methoden wie dem Lehrer-Schüler-Lernen besteht die Möglichkeit, dass sich Forschung und Entwicklungsmöglichkeiten von grossen Technologieunternehmen hin zu einer breiteren Forschungsgemeinschaft und mittelständischen Unternehmen verlagern können.

Für die Zukunft ist mit weiteren Durchbrüchen bei Modellgrösse und -leistung zu rechnen. Wichtig wird sein, die sozioökonomischen Implikationen, Risiken und ethischen Fragestellungen frühzeitig zu adressieren. Nur durch einen verantwortungsvollen Umgang kann das volle Potenzial dieser Technologien zum Wohle der Gesellschaft realisiert werden. Weitere interdisziplinäre Forschung ist erforderlich, um die Möglichkeiten und Grenzen generativer KI umfassend zu beleuchten und einen breiten gesellschaftlichen Dialog zu gestalten.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 8 (2024) — 10.18755/iw.2024.3