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

~~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