Analyse der Chancen und Risiken von KI-basierten Large Language Models am Beispiel von ChatGPT in der innerbetrieblichen IT

Cem Sidar Bulut

Hochschule Pforzheim
Tiefenbronner Straße 65
75175 Pforzheim
bulutcem@hs-pforzheim.de

Frank Morelli

Hochschule Pforzheim
Tiefenbronner Straße 65
75175 Pforzheim
frank.morelli@hs-pforzheim.de

SCHLÜSSELWÖRTER

Large Language Models, ChatGPT, Natural Language Processing

EINLEITUNG

Die digitale Transformation ermöglicht es zum einen, verschiedene KI-Technologien einzusetzen. Zum anderen erwachsen hieraus jedoch auch Risiken und Probleme. Bei einer der betroffenen Anwendungen handelt es sich um Chatbots unter Verwendung von künstlicher Intelligenz (KI). Chatbots sind vielseitig in unterschiedlichen Bereichen des Unternehmens einsetzbar. Ursprünglich waren diese primär dafür vorgesehen, einen menschlichen Dialog in Schriftform nachzustellen, um Kunden eine virtuelle und sofortige Unterstützung zu bieten. Der heutige Stand der Technik reicht weit darüber hinaus, wodurch sich neue Use Cases schaffen lassen.
Im Rahmen dieser Bachelorarbeit werden Möglichkeiten eines ChatGPT-Einsatzes in der innerbetrieblichen Informationstechnologie eines Unternehmens untersucht und diskutiert. Neben den Chancenpotenzialen fließen auch Risikoaspekte wie die Gefährdung der IT-Sicherheit oder der Umgang mit sensiblen Daten bei der Einbettung von ChatGPT ein. Aus dieser Betrachtungsweise heraus werden Handlungsempfehlungen abgeleitet.

ZIELSETZUNG

Der KI-basierte Chatbot-Markt lässt sich durch rapides Wachstum charakterisieren und erfreut sich starker Aufmerksamkeit in der Öffentlichkeit. Die dem Artikel zugrunde liegende Bachelorthesis verfolgt das generelle Ziel, für KI-basierten Chatbots, durch das Sammeln von möglichen Use Cases im Unternehmen, Transparenz für zugehörige EntscheiderInnen auf Basis des State-of-the Art zu schaffen. Ferner wird durch die Analyse von Chancen und Risiken im speziellen Fall von ChatGPT angestrebt, Handlungsempfehlungen für den innerbetrieblichen Einsatz im Unternehmen zu geben. Verantwortlichen in einer IT-Abteilung soll es darüber hinaus erleichtert werden, generische Einsatzmöglichkeiten von KI-basierten Chatbots auf das individuelle Unternehmensumfeld zu übertragen.

EINSATZMÖGLICHKEITEN

KI-basierte Algorithmen besitzen die Fähigkeit, eigenständig Schlussfolgerungen zu ziehen und aus dem Feedback neue Dinge zu erlernen. Konversationale KI beschreibt die Kommunikation mit Menschen mithilfe von natürlicher Sprache. Darunter fallen virtuelle Assistenten, konversationale Agenten und Chatbots. Im Rahmen von Natural Language Processing und insbesondere mit Large Language Models (LLMs) wird angestrebt, natürliche Sprache korrekt zu verarbeiten und in einem schlüssigen Kontext wieder auszugeben. Bei einer generativen KI ist es u.a. möglich, Antworten auf nicht im Vorfeld trainierten Fragen zu geben.
In diesem Kontext findet ChatGPT, ein Produkt der OpenAI, mit seinem Sprachmodell Davinci als generative NLP-getriebene Plattform besondere Aufmerksamkeit, sowohl generell in der Öffentlichkeit als auch speziell in der Unternehmenswelt. Es ist davon auszugehen, dass ChatGPT einen erheblichen Einfluss allgemein auf die zukünftige Arbeitsweise im Unternehmen und im Speziellen im IT-Bereich haben wird.
Bereits zum jetzigen State-of-the-Art gibt es viele Möglichkeiten, beim Einsatz von Routinearbeiten MitarbeiterInnen effizient zu unterstützen (vgl. Abb. 1).

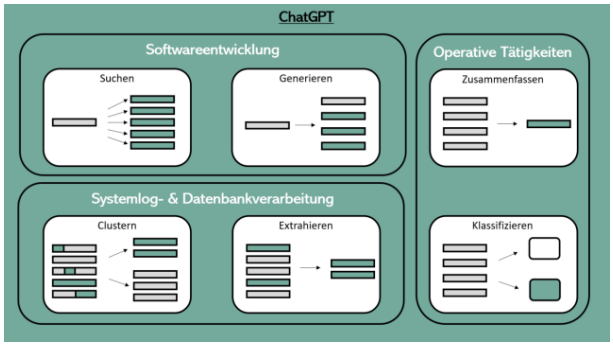

Abbildung 1: Darstellung der möglichen Einsatzpotentiale

Die Bewältigung von Aufgaben durch ChatGPT erfordert i.d.R. einen Impuls durch menschlich generierte Prompts. Entsprechend lässt sich der Einsatzbereich im Sinne einer virtuellen Assistenz charakterisieren.

RISIKEN BEIM EINSATZ

Eine Gefahr besteht darin, dass der Einsatz dieser Technologie kritisches Denken bei MitarbeiterInnen hemmt und die Bereitschaft, eigene innovative Lösungswege zu finden, abnimmt. Umgekehrt kann es durch die zur Verfügung gestellten Daten mit entsprechendem Prompting zu unbefugten Zugriffen und damit zu einer Verletzung der Datensicherheit kommen.
Bei ChatGPT wird seitens OpenAI ein Self Hosting nicht angeboten. Laut der offiziellen Privacy Policy von OpenAI werden persönliche Daten gesammelt. Entsprechend Artikel 6 der DSGVO erfordert dies i.d.R. die Einwilligung von allen Betroffenen. Generell ist im Sinne des Datenschutzes zu sagen, dass hierbei nur geringe Transparenz auf Anwendungsseite herrscht.
Bei ChatGPT ist eine selbstständige Arbeitsweise von ChatGPT zum heutigen Stand nicht möglich bzw. nicht ratsam, da falsche Schlussfolgerungen bzw. auf Annahmen basierende Outputs generiert werden. Im Sinne von Fairness ist darauf zu achten, dass es nicht zu einer Diskriminierung von Personengruppen kommt.
Mit dem Einsatz von ChatGPT ergeben sich auch juristische Problempunkte: Beispielsweise steigt mit dem ChatGPT Einsatz prinzipiell dass Risiko, dass durch externe Quellen, beispielsweise in Form von Phishing Kampagnen, Schaden im Unternehmen entsteht. Entsprechend stellt sich die Frage, wer in einem solchen Fall dafür zur Rechenschaft zu ziehen ist. Darüber hinaus erweisen sich urheberrechtliche Aspekte bei generativen KI-Modellen als ungeklärt.

HANDLUNGSEMPFEHLUNGEN

Generell müssen strikte Vorgaben gesetzt und eingehalten werden, um das Sprachmodell nachhaltig im Arbeitsalltag implementieren zu können. Ein überwachter Lernansatz ermöglicht es, eine Feinabstimmung für das LLM vorzunehmen und damit das Modell auf bestimmte Anwendungszenarien gezielt vorzubereiten. Bei ChatGTP stehen für die Optimierung des Sprachmodells drei Methoden zur Verfügung, Supervised Fine Tuning (SFT), Reward Model (RM)-Training und Reinforcement Learning via Proximal Policy Optimization (PPO). SFT eignet sich vor allem gut, wenn dem System hierfür bereits Trainingsdaten zur Verfügung stehen. Entsprechende Erkenntnisse wurden aus der Simulation im OpenAI Playground gewonnen: SFT trägt sowohl zur Begrenzung der mangelnden Informationssicherheit bei als auch zur Optimierung der Suche nach unternehmensinternen Informationen.
Eine weitere Handlungsempfehlung, um das Datenschutzrisiko zu verringern, ist der Aufbau von ChatGPT auf einem standardisierten Service wie z.B. dem von Azure OpenAI. Dies ermöglicht eine eigenständige Verwaltung des Sprachmodells. Weiterhin müssen flankierende Maßnahmen zur Sensibilisierung der Mitarbeiter-Innen im Hinblick auf Datensicherheit und Datenschutz getroffen werden.
Darüber hinaus kann man seit dem 25.04.2023 die Verarbeitung der Daten seitens OpenAI unterbinden. Eine Unternehmenspolicy sollte die zugehörige Einstellung zur Deaktivierung entsprechender Möglichkeiten standardmäßig vorschreiben.
Generell erweist sich eine regelmäßige Überwachung und Bewertung des Einsatzes von ChatGPT im Unternehmen als sinnvoll. Durch Kennzahlen ist es möglich, weitere Rückschlüsse auf Effektivität und Effizienz im laufenden Betrieb wie zum Beispiel dem Einsatz des Sprachmodells zu ziehen.

LITERATUR

Khowaja, Sunder Ali; Khuwaja, Parus; Dev, Kapal (2023): ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review. Online verfügbar unter http://arxiv.org/pdf/2305.03123v1.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N. et al. (2017): Attention Is All You Need. Online verfügbar unter http://arxiv.org/pdf/1706.03762v5.

Ouyang, Long; Wu, Jeff; Jiang, Xu; Almeida, Diogo; Wainwright, Carroll L.; Mishkin, Pamela et al. (2022): Training language models to follow instructions with human feedback. Online verfügbar unter http://arxiv.org/pdf/2203.02155v1.

Dong, Yihong; Jiang, Xue; Jin, Zhi; Li, Ge (2023): Self-collaboration Code Generation via ChatGPT. Online verfügbar unter http://arxiv.org/pdf/2304.07590v2.

Olmo, Alberto; Sreedharan, Sarath; Kambhampati, Subbarao (2021): GPT3-to-plan: Extracting plans from text using GPT-3. Online verfügbar unter http://arxiv.org/pdf/2106.07131v1.

Lim, Weng Marc; Gunasekara, Asanka; Pallant, Jessica Leigh; Pallant, Jason Ian; Pechenkina, Ekaterina (2023): Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. In: The International Journal of Management Education 21 (2), S. 100790. DOI: 10.1016/j.ijme.2023.100790.

Fuchs, Simon; Drieschner, Clemens; Wittges, Holger (2022): Improving Support Ticket Systems Using Machine Learning: A Literature Review. Online verfügbar unter https://scholarspace.manoa.hawaii.edu/handle/10125/79570.

Hartmann, Ernst A. (2022): Digitalisierung souverän gestalten II. Unter Mitarbeit von Ernst A. Hartmann. Berlin, Heidelberg: Springer Berlin Heidelberg. Online verfügbar unter https://library.oapen.org/handle/20.500.12657/51926

Rahaman, Md. Saidur; Ahsan, M. M. Tahmid; Anjum, Nishath; Rahman, Md. Mizanur; Rahman, Md Nafizur (2023): The AI Race is on! Google's Bard and Openai's Chatgpt Head to Head: An Opinion Article. In: SSRN Journal. DOI: 10.2139/ssrn.4351785.