VERBAND

DIGITALE ETHIK: UMGANG MIT SIEBEN FALLSTRICKEN BEI CHATGPT & CO.

Die Verbreitung von ChatGPT hat einen beispiellosen Hype rund um das Thema «generative KI» ausgelöst. Basierend auf möglichst präzisen Eingabeaufforderungen, auch «Prompts» genannt, lassen sich Texte, Bilder, Ton- und Videomaterial mit Hilfe von künstlicher Intelligenz (KI) erzeugen.

Wie jede Technologie bietet auch die generative KI nicht nur viele neue Möglichkeiten. Sie hat auch ihre Grenzen. Sie zu kennen ist wichtig, um Tools wie ChatGPT & Co. gewinnbringend und verantwortungsvoll einzusetzen. In diesem Merkblatt stellen wir sieben Fallstricke kurz vor, ergänzt um konkrete Use Cases. Ausserdem zeigen wir, wie sich potenzielle Risiken minimieren lassen, wobei der passende Lösungsansatz immer von der konkreten Situation abhängt.

1 Wahrheitsgehalt

Inhalte lassen sich mit Textgeneratoren mit wenig Aufwand und eloquent formuliert erstellen. Der Softwarehersteller BSI Business Systems Integration nutzt diese Möglichkeit, um neue Anwendungen zu bauen und zum Beispiel den Kundenservice mit automatisierten Textvorschlägen und Zusammenfassungen zu entlasten. Ein weiteres Beispiel ist die Marketingagentur Blue Glass. Sie nutzt Textgeneratoren für die Erstellung von Texten und Titelvorschlägen für Marketingkampagnen, das Identifizieren möglicher Nutzerbedürfnisse und das Clustering von Keyword-Listen. Die automatisch generierten Inhalte können aber auch falsch oder sogar erfunden sein, weil das Ergebnis solcher Textgeneratoren auf Wahrscheinlichkeiten und Zufall basiert, ausgehend von den Trainingsdaten.

> Zum Beispiel hat ein Team von Anwälten ChatGPT als Recherchetool eingesetzt. Die recherchierten Präzedenzfälle im Zusammenhang mit einem Rechtsstreit mit einer Airline waren schlicht erfunden. Das ChatGPT hat «halluziniert» (CNN 28.5.2023).

Lösungsansätze:

- Textgeneratoren nicht als Suchmaschine nutzen, weil sich der Mechanismus von jenem einer Suchmaschine grundlegend unterscheidet.
- Textgeneratoren nur für Themen nutzen, von denen der User eine gewisse Ahnung hat, um überprüfen zu können, ob die Ergebnisse wahr sind.
- Tools bevorzugen, die auf tatsächlich vorhandenen Daten basieren, damit es nicht oder möglichst wenig zu «Halluzinationen» (frei erfundenen Antworten) kommt.
- Tools bevorzugen, bei denen die verwendeten Quellen korrekt angegeben werden.
- Tools verwenden, die auf Open-Source basieren, um die Trainingsdaten zu kennen.
- Die Eingabeaufforderung (Prompt) möglichst genau formulieren.
- Inhalte vor ihrer Publikation auf ihren Wahrheitsgehalt überprüfen.

Swico
Lagerstrasse 33
CH-8004 Zürich

Tel. +41 44 446 90 90
www.swico.ch
info@swico.ch

SWICO
Der Wirtschaftsverband
für die digitale Schweiz

2 Vorurteile und Stereotypen

Die Inhalte von Text- und Bildgeneratoren basieren auf KI-Modellen, die mit enorm vielen Daten trainiert wurden. Im Zusammenspiel zwischen Mensch und Maschine entstehen dann neuartige Kreationen, zum Beispiel die Sujets der KI-Kampagne für das Weingut AI Mulinetto von Digitalkünstlerin Grit Wolany oder die Illustrationen für ein Kinderbuch vom Designer David Blum. Doch die Trainingsdaten von Text- und Bildgeneratoren können veraltet sein oder Verzerrungen enthalten, was sich im Ergebnis widerspiegelt.

Zum Beispiel werden vorhandene Stereotypen bezüglich Geschlecht und Ethnie beim Bildgenerator Stable Diffusion verstärkt. Frauen kommen selten vor, wenn Bilder über Ärzte, Anwälte oder Richter gesucht werden. Und Männer mit dunkler Haut begehen Verbrechen (Bloomberg 2023).

Lösungsansätze:

- Die Eingabeaufforderung (Prompt) möglichst genau formulieren und bestimmte Aspekte explizit in den Prompt integrieren (z.B. Sujet mit Menschen unterschiedlichen Alters).
- Die Text- und Bildvorschläge auf Vorurteile und Stereotypen überprüfen, zum Beispiel wenn eine Gruppe von Menschen sehr homogen zusammengesetzt ist.
- Tools verwenden, die auf Open-Source basieren, um die Trainingsdaten zu kennen.
- Für das Thema sensibilisiert sein, um potenzielle Diskriminierungen zu erkennen.

3 Inhaltsfilter

Sensible Themen oder beleidigende Witze sind bei vielen KI-Tools tabu. Ausserdem lassen sich automatisierte Inhalte auch für illegale oder unerwünschte Handlungen nutzen. Um dieses Missbrauchspotenzial zu reduzieren, können in den KI-Modellen inhaltliche Einschränkungen eingebaut werden. Zum Beispiel hat die Agentur Liip die Anwendung ZüriCityGPT entwickelt, um die Suchresultate für Zürcherinnen und Zürcher zu verbessern, basierend auf der Website der Stadt Zürich und weiteren Websites wie Zürich Tourismus. Bei unethischen Fragen wie Kaufmöglichkeiten für Drogen gibt der Chatbot keine Antwort. Beim Setzen von Inhaltsfiltern stellt sich immer die Frage nach ihrer Legitimation und Verhältnismässigkeit. Hinzu kommt, dass sich inhaltliche Einschränkungen umgehen lassen.

Zum Beispiel lassen sich die Inhaltsfilter von Chatbots umgehen, indem bestimmte Zeichenfolgen am Ende des Prompts hinzugefügt werden. Online sind zahlreiche weitere Methoden dokumentiert. Sie ermöglichen es, schädliche Inhalte wie Anleitungen zum Bombenbau und zum Ausspionieren von Nutzerdaten sogar automatisiert zu generieren (Netzwoche 31.7.2023).

Lösungsansätze:

- Den möglichen Missbrauch beim Einsatz eines KI-Tools von Beginn weg mitdenken, um im Krisenfall handlungsfähig zu sein (z.B. interne Zuständigkeit im Krisenfall).
- Filter mit Bedacht wählen, wenn sie entwickelt und eingesetzt werden sollen.

4 Täuschung

KI-Tools wie ChatGPT erlauben eine menschenähnliche Kommunikation, was die Nutzung für User sehr angenehm und einfach macht. Zum Beispiel hat Unic einen automatisierten Travel Planner für Schweiz Tourismus realisiert, der auf einer intelligenten Customer Journey basiert. Ziel ist es, das Kundenerlebnis zu verbessern, indem das System komplexe Suchanfragen interpretiert und personalisierte Reisevorschläge erstellt. Im Endausbau könnte ein Chatbot den Kundendienst von Schweiz Tourismus entlasten.

Swico
Lagerstrasse 33
CH-8004 Zürich

Tel. +41 44 446 90 90
www.swico.ch
info@swico.ch

SWICO
Der Wirtschaftsverband
für die digitale Schweiz

Wichtig ist, dass Menschen bei einem Chatbot wissen, dass sie mit einer Maschine und nicht mit einem Menschen interagieren. Das reduziert die Täuschungsgefahr und das Risiko, dass eine Vermenschlichung der Maschine stattfindet.

> Zum Beispiel bietet das Start-up Koko über die Chatplattform Discord eine Peer-to-Peer-Unterstützung für Menschen in psychischen Krisen und Ratsuchende an. Als die Hilfesuchenden erfuhren, dass gewisse Nachrichten mit ChatGPT verfasst wurden, reagierten sie verstört. Diese «simulierte Empathie» widersprach ihrer Erwartung und der Kommunikation von Koko, von einem Menschen unterstützt zu werden. (Vice 10.1.2023).

Lösungsansätze:

- Bei der Gestaltung der Kundenschnittstelle darauf achten, dass die Kommunikation nicht zu menschenähnlich ist (z.B. bei den Antworten auf Emojis verzichten, mit einem Disclaimer klarstellen, dass die generierte Information falsch sein kann).
- Beim Einsatz von Chatbots sicherstellen, dass die Nutzer wissen, dass sie nicht mit einem Menschen interagieren, sondern mit einer Maschine.
- Insbesondere in sensiblen Bereichen überprüfen, ob die Interaktion mit einem Chatbot den Erwartungen der User entspricht.

5 Manipulation

Mit Hilfe von KI-Tools können Texte, Bilder, Ton- und Videomaterial mit wenig Aufwand und in echt wirkender Qualität erzeugt werden. Zum Beispiel lassen sich fiktive Sujets für Kampagnen oder Avatare für Verkaufsgespräche und Erklärvideos automatisiert erstellen. Für Aufsehen sorgte der Schweizer Privatsender M Le Média: Seit April 2023 wird das Wetter von einer Avatar-Moderatorin präsentiert. Problematisch ist, wenn die neuen Möglichkeiten zum Zweck der Täuschung oder zur Manipulation der öffentlichen Meinung eingesetzt werden. Dies kann beispielsweise der Fall sein, wenn Personen des öffentlichen Lebens Aussagen in den Mund gelegt werden, die sie nie machen würden. Solche Desinformationen verbreiten sich schnell und sind eine Gefahr für Demokratien.

> Zum Beispiel haben Unbekannte im Netz ein Deepfake-Video verbreitet, in dem der ukrainische Präsident Wolodymyr Selenskyj vermeintlich zur Kapitulation aufruft (zdf heute 18.3.2022).

Lösungsansätze:

- Bei fiktiven Sujets sicherstellen, dass sie nicht mit einer realen Situation verwechselt werden. Bei einer Verwechslungsgefahr das Sujet überdenken.
- Bei KI-generierten Inhalten deklarieren, dass sie mit KI erzeugt wurden.
- Für das Thema sensibilisiert sein, um keine Desinformation zu produzieren oder über die eigenen Kanäle zu verbreiten.

6 Datenschutz und Datensicherheit

Die in einem Textgenerator eingegebenen Daten werden gespeichert, um auf die gestellte Frage eine Antwort geben zu können. Zum Beispiel nutzt die Helvetia Versicherung den ChatGPT-basierten Chatbot Clara, um Kundenfragen auf der Website zu beantworten. Dabei wird empfohlen, keine personenbezogenen Daten in den Chatverlauf einzugeben, wenn man sicherstellen möchte, dass diese Daten nicht weitergegeben werden. In Italien wurde ChatGPT im Frühling 2023 sogar kurzzeitig verboten aufgrund möglicher Verstösse gegen die Datenschutz-Grundverordnung (DSGVO) in der EU. Ob dies tatsächlich der Fall ist, wird derzeit von den europäischen Datenschutzbehörden geklärt.

Swico
Lagerstrasse 33
CH-8004 Zürich

Tel. +41 44 446 90 90
www.swico.ch
info@swico.ch

SWICO
Der Wirtschaftsverband
für die digitale Schweiz

Vorsicht ist auch geboten mit Blick auf Unternehmensgeheimnisse: Wenn Mitarbeitende sensible Unternehmensdaten eingeben, zum Beispiel für die Zusammenfassung von Sitzungsprotokollen oder um Programmcode zu generieren, können diese Daten für das Training der KI-Modelle verwendet werden und die Informationen in den Ergebnissen für andere Nutzer landen. Sensible Daten können ausserdem in falsche Hände gelangen und missbraucht werden, etwa um Identitäten zu stehlen. Sogenannte «Prompt-Injections»-Angriffe könnten die KI-Tools sogar dazu bringen, Informationen preiszugeben, die sie nicht preisgeben sollten.

Auf den Darknet-Marktplätzen sind bereits viele Zugangsdaten für ChatGPT aufgetaucht. Samsung hat aus diesem Grund eine eigene, sicherere GPT-Instanz eingekauft und Google-Engineers dürfen den hauseigenen Chatbot Bard aus Sicherheitsgründen nicht einsetzen. (Inside-IT 21.6.2023)

Lösungsansätze

- Prüfen, bei welchen Tätigkeiten kein Textgenerator eingesetzt werden sollte.
- Die Datenschutzrichtlinien der Anbieter von KI-Tools lesen, um zu wissen, wie die Daten verwendet werden.
- Bei der Eingabeaufforderung (Prompt) keine vertraulichen, sensiblen oder persönlichen Daten eingeben, weil diese in fremde Hände gelangen können.
- Für das Thema sensibilisiert sein, um sich den Herausforderungen beim Datenschutz und bei der Datensicherheit bewusst zu sein.

7 Urheberrecht

Die mit KI generierten Inhalte fallen nicht unter das Urheberrecht, weil gemäss aktueller Rechtsprechung nur Menschen Schöpfer neuer Werke sein können. Dies bedeutet im Umkehrschluss, dass KI-Bilder nicht geschützt werden können. Gleichzeitig werden die KI-Modelle dieser Tools mit umfangreichen Daten trainiert, die zum Teil urheberrechtlich geschützt sind. Es gibt sogar Tools, die Werke im Stile eines bestimmten Künstlers produzieren können. Ob diese Fälle eine Verletzung des Urheberrechts darstellen oder unter die Fair-Use-Doktrin fallen, ist derzeit unklar.

Aufgrund rechtlicher Bedenken haben Getty Images und Shutterstock bereits reagiert: Sie haben KI-Bilder auf ihren Plattformen entfernt und den Verkauf verboten. Die Klärung der rechtlichen Situation ist in Gange, etwa im Rahmen des AI Acts der EU sowie über hängige Gerichtsfälle in den USA.

Zum Beispiel werfen die US-Komikerin Sarah Silverman und die Autoren Christopher Golden und Richard Kadrey dem Unternehmen OpenAI vor, ihre urheberrechtlich geschützten Werke ohne Einwilligung zum Trainieren ihrer KI-Modelle für ChatGPT verwendet zu haben. Dagegen gehen sie gerichtlich vor (SRF 11.7.2023).

Lösungsansätze

- KI-Tools bevorzugen, bei denen die verwendeten Quellen korrekt angegeben werden.
- Verzicht auf KI-Bilder, wenn eine Exklusivität wichtig ist, da diese nicht möglich ist.
- Das Einverständnis einholen, wenn Daten oder Inhalte von Dritten genutzt werden.
- Veröffentlichung von Zusammenfassungen urheberrechtlich geschützter Daten, die für das Training von KI-Modellen verwendet wurden.

Swico
Lagerstrasse 33
CH-8004 Zürich

Tel. +41 44 446 90 90
www.swico.ch
info@swico.ch

Für Rückfragen:

Judith Bellaiche
SWICO
Mobile: +41 79 2175645

Swico
Lagerstrasse 33
CH-8004 Zürich

Tel. +41 44 446 90 90
www.swico.ch
info@swico.ch