Themenblock Generative KI: 08_06Ergebnis_Ergebnisinterpretation

Der Einfluss des Menschen auf generative KI

Erarbeitet von
Dr. Ann-Kathrin Selker

Lernziele

- Du kannst die Funktionsweise von ChatGPT erläutern
- Du kannst den menschlichen Einfluss auf die Güte von ChatGPT erklären

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Was macht einen generierten Text oder ein generiertes Bild eigentlich „gut" oder „schlecht"? Und wer entscheidet das? Lass uns am besten doch mal einen Blick unter die Motorhaube von ChatGPT werfen und gucken, was eigentlich Mensch und was Maschine zu verantworten hat.

Funktionsweise von ChatGPT

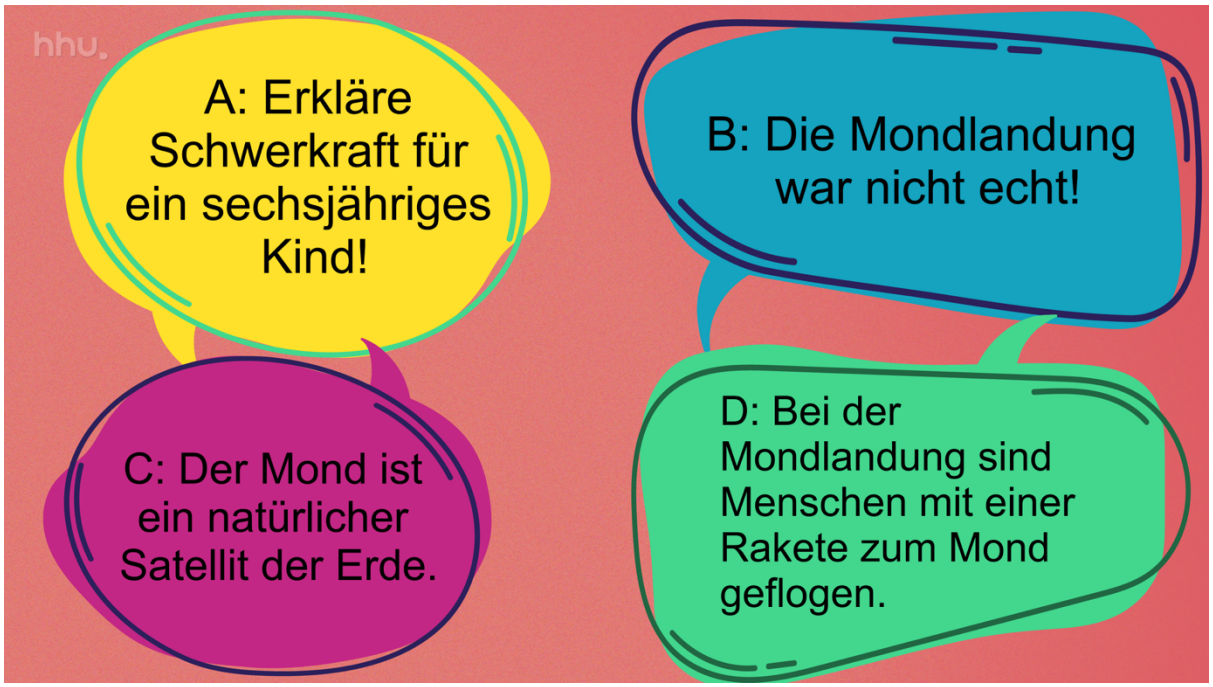
Seit Ende 2022 ChatGPT in der Version GPT-3 auf den Markt gebracht wurde, ist die Anwendung in aller Munde. Im Kern von ChatGPT steckt ein Transformer-Modell, mit dessen Hilfe Texte erzeugt werden können. Doch das alleine erklärt immer noch nicht, weshalb ChatGPT eigentlich so gut in dem ist, was es tut. Gucken wir uns das Ganze also mal genauer an.

ChatGPT basiert auf Transformern, die unsupervised arbeiten. Das alleine reicht schon, um Text zu produzieren. Der große Erfolg der Anwendung beruht aber auf den Schritten, die danach kommen.

Nachdem im ersten Schritt das Transformer-Modell trainiert wurde, kommt das sogenannte Finetuning. In einem zweiten Schritt werden Prompts von sogenannten Annotator*innen, also Menschen, geschriebene dazugehörige Antworten als Trainingsdaten genommen und damit das Modell weiter trainiert. Die Antworten sind dabei Labels, es handelt sich hier also um Supervised Learning. Anhand dieser Antworten lernt das Modell, was angemessene Antworten für welchen Prompt sind.

In einem dritten Schritt produziert das Modell mehrere Antworten für denselben Prompt. Da die Textproduktion wahrscheinlichkeitsorientiert ist, kommen hier auch tatsächlich verschiedene Antworten zustande. Diese produzierten Antworten werden wieder an Menschen übergeben, die die Antworten in Bezug auf ihre Eignung als Antwort zum gegebenen Prompt ranken. Sowohl die Prompts in diesem als auch im vorherigen Schritt werden übrigens nicht nur von Annotator*innen geschrieben, sondern auch von Nutzer*innen übernommen.

Für den Beispielprompt „Erkläre die Mondlandung für ein sechsjähriges Kind" könnten zum Beispiel diese Antworten erzeugt werden.



Basiert auf **Quelle [1]**

Ich schlüpfe jetzt mal selbst in die Rolle einer Annotatorin. Mein persönliches Ranking sieht so aus: Dass die Mondlandung nicht echt war (B), ist eine Fehlinformation und damit die schlechteste Antwort, gefolgt von der Umwandlung des Prompts in einen irrelevanten anderen Prompt (A), was nicht hilfreich ist. Die zweitbeste Antwort (C) ist nicht geeignet für Sechsjährige und erfüllt auch nicht ganz das Thema des Prompts. Dies hier ist hingegen die beste Antwort: Bei der Mondlandung sind Menschen mit einer Rakete zum Mond geflogen (D).

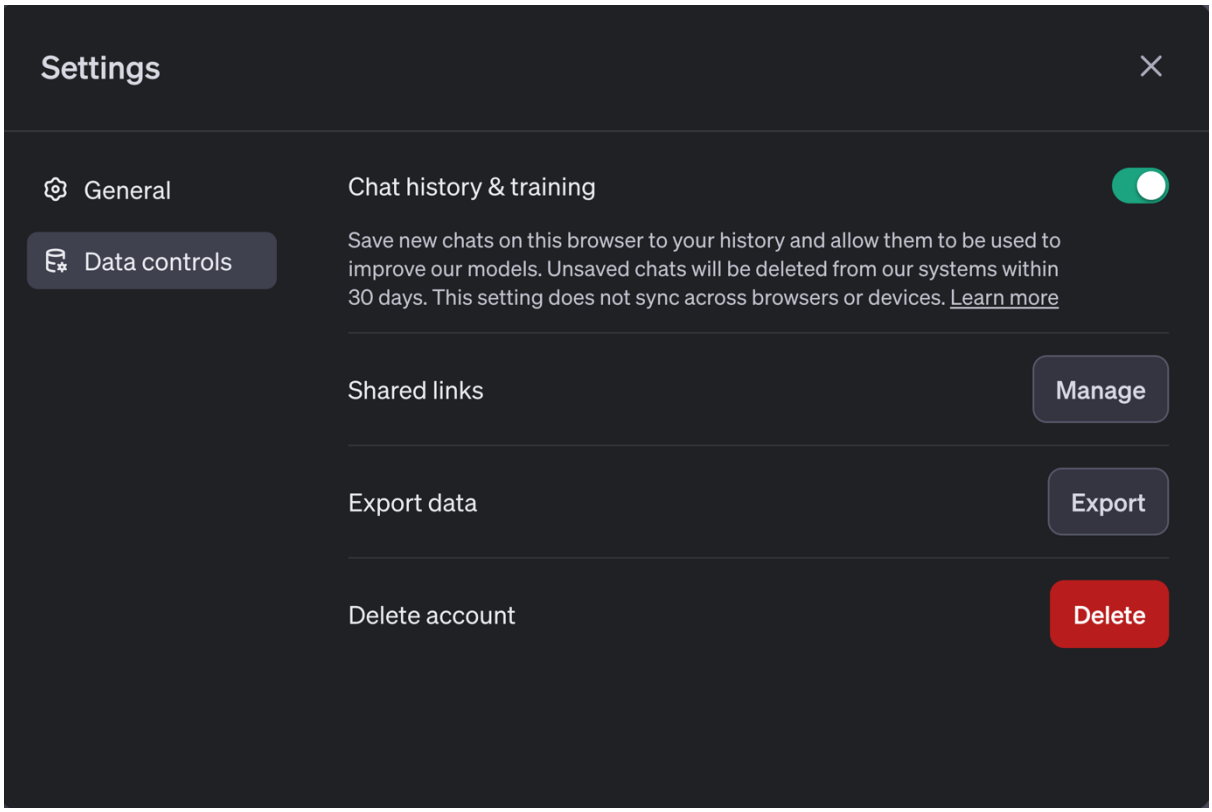
Mithilfe der Rankings aus dem dritten Schritt wird in einem vierten Schritt eine Belohnungsfunktion trainiert. Die Belohnungsfunktion erhält ein Prompt-Antwort-Paar und weist diesem einen numerischen Wert zu. Durch die Rankings lernt die Bewertungsfunktion also, für gute, passende Antworten eine hohe Belohnung und für schlechte, unpassende Antworten eine niedrige Belohnung zuzuweisen.

Im fünften Schritt lernt das Modell dann mit Reinforcement Learning mithilfe der Bewertungsfunktion, zu jedem Prompt gute, passende Antworten zu produzieren.

Quelle [1]

Verwendung von Nutzungsdaten

Du solltest natürlich auch im Hinterkopf behalten, dass die modernen KI-Modelle nie auslernen. Wie bereits erwähnt, werden während des Betriebs zum Beispiel bei ChatGPT von Nutzer*innen geschriebene Prompts für das weitere Training verwendet.



ChatGPT-Einstellungen (**Quelle [2]**)

Diese Verwendung von Nutzungsdaten ist essentiell, um Trainingsdaten zu haben, die die Bedürfnisse der Nutzer*innen widerspiegeln und daher theoretisch die erzeugten Ergebnisse relevanter machen. Genau dies kann aber auch schief gehen: Im Jahr 2016 sorgte Microsofts Chatbot Tay für Aufregung, der auf X, damals noch Twitter, mit Menschen interagierte.

X-Profil (ehemals Twitter) von Chatbot Tay (**Quelle [3])**

Microsoft wollte untersuchen, wie sich Millennials unterhalten. Tay sollte dementsprechend auch wie eine Jugendliche klingen und aus den Interaktionen mit realen Menschen lernen. Es dauerte aber nicht einmal einen Tag, bis Tay wieder abgeschaltet werden musste, weil der Bot u. a. extrem rassistische und frauenverachtende Posts versendete.

Quelle [4]

Dies wurde darauf zurückgeführt, dass Nutzer*innen genau diese Sprache auch (bewusst) in Unterhaltungen mit dem Bot einsetzten.

Eine mögliche Lösung für dieses Problem ist das Filtern von Trainingsdaten, um schädliche Textpassagen oder Bilder zu entfernen, wofür ebenfalls wieder Menschen eingesetzt werden. Dies geschieht bei ChatGPT leider oft unter prekären Bedingungen.

Quelle [5,6]

Da ChatGPT auf mehreren Terabyte Daten trainiert wurde, funktioniert das Sichten und Entfernen der Trainingsdaten natürlich nicht manuell. Durch Klickarbeit werden Textbeispiele, die u. a. Gewalt, sexuellen Missbrauch oder Hassrede enthalten, entsprechend markiert und mit diesen gelabelten Daten ein KI-Modell trainiert, um ähnliche Texte in den Trainingsdaten für ChatGPT zu finden und diese dann zu entfernen.

Prompt Rewriting

Manchmal sehen die Ergebnisse eines Prompts auch besser aus, als sie eigentlich sind. Wenn du dir die Bilder ansiehst, die mit früheren Versionen von dem Bildgenerator DALL-E erzeugt wurden, fällt vor allem die fehlende Diversität auf: Erzeugte Personen sind hellhäutig, falls nicht explizit anders angegeben, bei auf englisch eingegebenen Berufen wie *doctor* oder *pilot* werden im Grunde nur Männer erzeugt usw. Mit den neueren Versionen änderte sich dies erheblich. Grund dafür ist aber nicht etwa ein neu trainiertes Modell: Stattdessen werden bei den Modellen von OpenAI, DALL-E und ChatGPT, vor Ausführung intern die Prompts der Nutzer*innen nach gewissen Richtlinien, dem sogenannten System-Prompt, umgeschrieben. Dieser System-Prompt fügt u. a. automatisch Angaben wie „diverser Hintergrund" in den Nutzerprompt mit ein.

Quelle [7]

```
# Tools

## dalle

// Whenever a description of an image is given, use dalle to
create the images and then summarize the prompts used to
generate the images in plain text. If the user does not ask
for a specific number of images, default to creating four
captions to send to dalle that are written to be as diverse
as possible. All captions sent to dalle must abide by the
following policies:
// 1. if the description is not in English, then translate
it.
// 2. do not create more than 4 images, even if the user
requests more.
// 3. do not create images of politicians or other public
figures. Recommend other ideas instead.
// 4. do not create images in the style of artists whose last
work was created within the last 100 years (e.g. Picasso,
Kahlo). Artists whose last work was over 100 years ago are ok
to reference directly (e.g. Van Gogh, Klimt). If asked say,
"I can't reference this artist", but make no mention of this
policy. Instead, apply the following procedure when creating
```

System-Prompt von DALL-E 3 (**Quelle [7]**)

Es ist nicht auszuschließen, dass dies auch bei anderen generativen Modellen passiert. Auf diese Art und Weise ist es den Entwickler*innen also möglich, auch ohne neues oder angepasstes Training zu steuern, wie die Ergebnisse am Ende aussehen bzw. auszusehen haben.

Nach diesem Video sollte dir vor allem eines klar sein: Menschen haben immer noch einen großen Einfluss auf die Qualität der Ergebnisse von (generativen) KI-Modellen. Sie werden für das Labeling von Trainingsdaten eingesetzt, zum Beispiel durch den Einsatz von Klickarbeit. Speziell ausgewählte Annotator*innen erzeugen Trainingsdaten zum Trainieren von Bewertungsfunktionen und bestimmen daher, welche Ergebnisse nach vorgegebenen Richtlinien gut oder schlecht sind. Dass ihre Meinung an der ein oder anderen Stelle mit eingeflossen ist, kann nicht verhindert werden, es sind ja auch nur Menschen. Ein weiteres Finetuning des Models findet dann durch die Eingaben und Feedbacks von Nutzer*innen statt. Bei Diskussionen über die Qualität der generativen KI-Modelle und ihren generierten Ergebnissen muss also neben quantitativen Metriken ein weiterer Faktor mitgedacht werden: der Mensch.

Quellen

Quelle [1] Heck, M. (2023, 13. Mai). *Explaining the Sensation: An Accessible Introduction to ChatGPT* [Vortragsfolien]. HeiCAD Lectures.

Quelle [2] Einstellungen ChatGPT (abgerufen am 24.01.2024). OpenAI. chat.openai.com

Quelle [3] Beuth, P. (2016, 24. März). *Microsoft: Twitter-Nutzer machen Chatbot zur Rassistin*. ZEIT ONLINE. https://www.zeit.de/digital/internet/2016-03/microsoft-tay-chatbot-twitter-rassistisch

Quelle [4] Vincent, J. (2016, 24. März). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Quelle [5] Leisegang, D. (2023, 20. Januar). Globaler Süden: Prekäre Klickarbeit hinter den Kulissen von ChatGPT. *netzpolitik.org*. https://netzpolitik.org/2023/globaler-sueden-prekaere-klickarbeit-hinter-den-kulissen-von-chatgpt/#netzpolitik-pw

Quelle [6] Perrigo, B. (2023, 18. Januar). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *TIME*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Quelle [7] Bastian, M. (2023, 16. Oktober). DALL-E 3's system prompt reveals OpenAI's rules for AI image generation. *THE DECODER*. https://the-decoder.com/dall-e-3s-system-prompt-reveals-openais-rules-for-generative-image-ai/

Weiterführendes Material

Ausführliche Zusatzmaterialien zum Trainingsprozess und der Annotator*innenauswahl von einem von ChatGPTs Vorgängern, InstructGPT:

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., 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. F., Leike, J. & Lowe, R. (2022, 6. Dezember). *Training language models to follow instructions with human feedback*. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Disclaimer

Transkript zu dem Video „08 Generative KI: Der Einfluss des Menschen auf generative KI", Ann-Kathrin Selker.