KULTUR NEU ENTDECKEN

SWR KULTUR

Das Wissen

Wie KI bessere Antworten liefert – Prompt Engineering

Von Aeneas Rooch

Sendung vom: Dienstag, 12. November 2024, 8.30 Uhr
Redaktion: Charlotte Grieser
Autorenproduktion
Produktion: SWR 2024

KI-Chatbots wie ChatGPT liefern bessere Ergebnisse, wenn wir besser fragen. Die manchmal kuriosen Techniken des Prompt Engineering zielen auf die internen Mechanismen der Algorithmen.

MANUSKRIPT

Musikakzent

Autor:
Viele Menschen nutzen künstliche Intelligenz – für die Arbeit, für die Schule oder privat. Sie verwenden KI-Chatbots wie ChatGPT, Gemini oder Copilot, um lange Artikel zusammenzufassen, um E-Mails zu schreiben, Texte umzuformulieren, Ideen zu sammeln und Vieles mehr.

Die Künstliche Intelligenz kann erstaunlich viel. Aber wir können noch mehr herausholen: mit Prompt Engineering.

Ansage:
Wie KI bessere Antworten liefert – Prompt Engineering. Von Aeneas Rooch.

Musikakzent

Autor:
Wenn wir unsere Anweisungen an einen KI-Chatbot, die sogenannten Prompts, auf bestimmte Arten formulieren, wenn wir also unsere Beschreibung, was der KI-Chatbot für uns tun soll, konstruieren, dann bekommen wir ausführlichere, kreativere und richtigere Antworten.

Manche Techniken bei diesem Prompt Engineering klingen plausibel, andere kurios. Ich werde sie ausprobieren. Und ich werde der Frage nachgehen: Wieso funktioniert Prompt Engineering überhaupt? Wie entsteht eine Antwort tief im Inneren von ChatGPT und Co. – und wie kann die Art, wie ich meine Frage stelle, daran etwas verbessern?

Atmo: Tastatur

Autor:
Ich lasse ChatGPT eine lange, unübersichtliche E-Mail zusammenfassen. Sie enthält Ideen und Absprachen für ein Buchprojekt, zwischen meinem Agenten und mir.

Prompt:
Fasse mir diesen Text zusammen.

Computerstimme:
Die Illustratorin aus Wien ist interessiert …

Autor:
Der Chatbot fasst die lange E-Mail richtig zusammen. Erstaunlich. Er ist nur ein Computerprogramm, aber er scheint zu verstehen, worum es in Mail geht und was darin die relevanten Informationen sind. Die Zusammenfassung geht aber noch besser.

Es hat sich herausgestellt, dass man von KI-Chatbots wie ChatGPT passendere Antworten erhält, wenn man bei der Anfrage bestimmte Dinge sagt. Was immer wirkt:

Beschreibe der KI präzise, wie die Antwort aussehen soll: Soll ein Gedicht herauskommen, ein Post für Social Media, eine E-Mail? Wozu wird es gebraucht? Welcher Tonfall wird gewünscht – lustig, geschäftlich, salopp? Soll nur ein einziger Satz geschrieben werden oder eine halbe Seite? Was auch wichtig ist: Gib der KI eine Rolle. Ist sie jetzt gerade Lehrerin, Expertin für Marketing, Notarin, Schriftstellerin? Ich probiere es aus.

Atmo: Tastatur

Prompt:
Agiere als Verlagskauffrau mit großem Organisationstalent. Fasse diese Notizen in einem einzigen Absatz zusammen. Erstelle danach eine Liste der genannten Personen und zähle jeweils die wichtigsten Punkte auf, die sie gesagt haben. Schreibe zum Schluss eine Liste mit den vereinbarten nächsten Schritten.

Computerstimme:
Zusammenfassung der Email: Der Autor hat Kontakt zur …

Autor:
KI braucht Kontext. Das ist wie bei uns Menschen, sagt Doris Weßels. Sie ist Professorin für Wirtschaftsinformatik an der Fachhochschule Kiel.

O-Ton Doris Weßels, Wirtschaftsinformatikerin:
Wenn ich jetzt, ohne dass ich Ihnen Kontext-Informationen gebe, ohne Hintergrundwissen, ohne eine vernünftige Erklärung, dann können Sie auch gar keine gute Leistung bringen. Ich auch nicht. Das heißt, wir brauchen noch immer Hintergrundinformationen, auch als Menschen. Und so ähnlich ist es auch mit diesem KI-System.

Autor:
Je weniger die KI raten muss, was ich von ihr will, desto wahrscheinlicher ist es, dass ich es auch bekomme. Also: Kontext geben und Details nennen – das ist bereits einfaches Prompt Engineering. Einfache Eingabe-Technik. Und wenn dabei nicht das herauskommt, was man gerne hätte, kann man nachbessern.

Atmo: Tastatur

Prompt:
Schreibe mir die Liste von eben als Gedicht.

Computerstimme:
Der Autor sprach: Ich fand die Frau, die Illustratorin …

Autor:
Das System versteht, was das ist – „die Liste von eben“. Und es versteht, was ein Gedicht ausmacht, es findet Wörter, die sich reimen, und es setzt sie passend zusammen. Auch das ist Prompt Engineering: Sich nicht mit dem ersten Wurf eines KI-Systems zufriedengeben, sondern das Ergebnis schrittweise verbessern. Doris Weßels rät: einfach drauf los probieren, wie im Gespräch mit einem echten Menschen:

O-Ton Doris Weßels:
Man gibt in ganz natürlicher Sprache ein, was man gerne hätte, so wie ich auch bei Ihnen das jetzt sagen würde, was ich ganz gerne von Ihnen hätte. Und wenn ich feststelle, Sie liefern etwas ab, was nicht meinen Vorstellungen genügt, dann hake ich noch mal nach, dann erkläre ich es in einem zweiten Versuch, und so nähern wir uns iterativ dann an und irgendwann bekomme ich auch das Ergebnis.

Autor:
Wenn der Text zu lang war, schreibe ich:

Prompt:
Antworte kürzer.

Autor:
Wenn mir der Stil nicht gefällt, sage ich konkret, welchen Tonfall ich wünsche:

Prompt:
Schreibe es witziger.

Autor:
Wenn die Antwort zu kompliziert ist, bestelle ich:

Prompt:
Erkläre es mir, als wäre ich 10 Jahre alt.

Autor:
Exakt benennen, was man möchte. Schrittweise verbessern. Auf diese Techniken kommt man womöglich auch von selbst, beim Rumprobieren mit Claude, Gemini und Co. Andere Techniken beim Prompt Engineering liegen nicht so auf der Hand, obwohl auch sie durchaus natürlich klingen:

O-Ton Doris Weßels:
Eine Hilfestellung, die wir auch in menschlicher Kommunikation häufig geben, wenn wir so Kinder betreuen oder Lehrer machen: dass wir dazu anregen, Schritt für Schritt etwas zu durchlaufen, nicht gleich drauf loszurennen, sondern ein bisschen zu überlegen und Schritt für Schritt etwas zu machen. Und im Bereich Prompting nennt man das Chain-of-Thought- Prompting, also Gedankenketten-Prompting. Das bedeutet, wenn man einem solchen System sagt: Bitte Schritt für Schritt, so gibt man ihm mit solch einer Gedankenkette, die man dann produzieren lässt, auch Hilfestellung.

Atmo: Katze

Autor:
Wenn ich ChatGPT auffordere:

Prompt:
Ich habe eine rötlich-getigerte Katze. Gib mir fünf Ideen für Namen.

Autor:
Dann bekomme ich irgendwas, das mehr oder weniger passt.

Computerstimme:
Hier sind fünf Namensideen für Deine rötlich getigerte Katze …

Autor:
Das System kann passender antworten, wenn ich ihm Hilfestellung gebe, durch „Chain of Thought Prompting“, wenn ich es also zwinge, gedankliche Schritte zu vollziehen.

Atmo: Tastatur

Prompt:
Gib mir fünf Namen für eine rötlich-getigerte Katze. Begründe, warum du diese Namen ausgewählt hast. Überlegen wir schrittweise, um geeignete Namen zu finden.

Computerstimme:
Lass uns schrittweise überlegen, wie wir passende Namen …

Autor:
Wenn das Modell angehalten wird, Gedankenschritte zu vollziehen, wird es besser. Dabei ist das, was im Algorithmus abläuft, natürlich kein Gedanke.

O-Ton Doris Weßels:
Zumindest nicht das, was wir als menschlichen Gedankengang bewerten. Es ist so eine Fassade, die hier aufgebaut wird, die uns glauben lässt, dass dort wirklich Schritt für Schritt nachgedacht wurde, so wie wir Menschen nachdenken. Und das ist es natürlich nicht, sondern es ist ein maschineller Prozess, der dort abgelaufen ist.

Autor:
Die KI ist ein Computerprogramm. Sie berechnet, welche Wörter in welcher Reihenfolge am wahrscheinlichsten zu der gestellten Aufgabe passen. Diese statistischen Berechnungen laufen auf so vielen Ebenen ab und beziehen so viele Einflüsse mit ein, dass wir die Berechnungen nicht mehr im Detail nachvollziehen können. Wenn wir das Programm auffordern nachzudenken, mag es wirken, als machte es sich wirklich Gedanken – in seinem Inneren läuft aber nur ein weiterer Rechenprozess ab.

O-Ton Doris Weßels:
Der aber, und das versöhnt uns dann mit dem Ergebnis, der aber bessere Ergebnisse erzielt, als wenn wir das nicht gemacht hätten. Ja, auch das ist befremdlich, das irritiert auch, aber es funktioniert.

Musikakzent

Autor:
ChatGPT wurden Abermillionen von Texten vorgesetzt: Bücher, Zeitungsartikel, Webseiten, Kommentare aus sozialen Netzwerken. In diesem Material hat das Programm nach Regelmäßigkeiten gesucht, wie: Welche Silben folgen aufeinander?

Welche Wörter folgen aufeinander? Anhand der Muster, die das Programm in den Beispielen erkannt hat, soll es dann unbekannte Texte, die ihm vorgesetzt werden, ergänzen: Es soll jeweils ein Wort anhängen, passend dazu, wie es das in den Beispielen gesehen hat. Das erklärt der Informatiker Dr. Thomas Arnold von der Technischen Universität Darmstadt:

O-Ton Thomas Arnold, Informatiker:
Text geht rein, wird mit sehr komplizierten Wahrscheinlichkeitsberechnungen und Verteilungen verarbeitet und raus kommt im Prinzip wieder eine Wahrscheinlichkeitsverteilung: Was könnte denn das nächste Wort sein? Und das wird dann hinten drangehängt an das, was man eingegeben hat. Und so wird immer ein Wort nach dem anderen prozessiert und weiter ausgegeben.

Autor:
Wie geht zum Beispiel dieser der Satz weiter?

Prompt:
Ich wünsche Ihnen einen angenehmen… Punkt, Punkt, Punkt.

Autor:
In den Beispieltexten hat das Modell mehrere Möglichkeiten gefunden.

O-Ton Thomas Arnold:
Ich wünsche Ihnen einen angenehmen Abend, einen angenehmen Morgen, einen angenehmen Flug. Und so würde dann einfach eines von diesen Möglichkeiten ausgewählt werden.

Autor:
Kombinationen, die in den Beispieltexten oft vorkamen, verwendet das Modell entsprechend öfter. So arbeiten große Sprachmodelle, sogenannte Large Language Models, schon lange: mit riesigen Tabellen, welche Wörter in welcher Kombination wie oft vorkommen. Bis 2017: Da kam der Durchbruch, mit einer neuen Softwarearchitektur, einem sogenannten Transformer-Modell. Daher kommt der Name ChatGPT: Chat Generative Pre-trained Transformer.

Musikakzent

Autor:
Das Modell besitzt nicht nur Listen, welche Wörter in welcher Reihenfolge wie häufig vorkommen. Sondern es weiß auch, welche Wörter inhaltlich zusammengehören oder ähnlich sind. Damit kann es flexibel reagieren, wenn es einem Ausdruck begegnet, den es aus seinem Trainingsmaterial nicht genau so kennt.

Nehmen wir zum Beispiel mal meinen Namen: Aeneas Rooch. Der ist selten. Es kann gut sein, dass der Name in den Texten, mit denen die KI trainiert wurde, einfach nicht vorkam. Ein altes Modell hat jetzt ein Problem, wenn es den Satz vervollständigen soll:

Prompt:
Aeneas Rooch ist der… Punkt, Punkt, Punkt.

O-Ton Thomas Arnold:
Ein herkömmliches Sprachmodell würde jetzt hier einfach versagen. Also würde sagen: Okay, ‚Aeneas Rooch ist' habe ich nie gesehen, ich habe keine Ahnung, wie ich es fortsetzen soll, und würde entweder gar nichts tun oder würde einfach ein völlig zufälliges Wort auswählen, das gar nichts mit diesem Satz zu tun hätte.

Autor:
Tranformer-Modelle wie ChatGPT jedoch können abstrahieren, erläutert Thomas Arnold:

O-Ton Thomas Arnold:
Okay, Aeneas ist ein Vorname, das habe ich schon mal gesehen, zwar nicht in genau dieser Konstellation, aber ich weiß, dass es irgendwie ein Vorname ist. Rooch scheint wohl irgendwie ein Nachname zu sein, kommt auch nach einem Vornamen. Das heißt, es kommt einfach ‚irgendein Name ist der…' und dann kann ich fortsetzen: ‚Aeneas Rooch ist der beste…' und dann so weiter, dann kann ich eben trotzdem irgendwie einen diesen Satz vervollständigen, auch wenn ich genau diese Wortkonstellation noch nie gesehen habe.

Musikakzent

Autor:
Auf die gleiche Art hat es auch gelernt, mit hoher Wahrscheinlichkeit grammatikalisch korrekte Sätze zu bilden – ohne eine einzige Regel zu kennen, wie Sprache funktioniert.

O-Ton Thomas Arnold:
Auch das ist einfach durch Häufigkeiten gelernt. In den Trainingsdaten, die dafür hergenommen wurden, um ChatGPT zu trainieren, wurde ja im Prinzip einfach das Internet gecrawlt, so viel wie möglich Daten abgegriffen, alles, was frei verfügbar ist und manchmal vielleicht auch nicht frei verfügbar ist, so genau weiß man das ja leider nicht.

Autor:
Die meisten Firmen verraten nicht, welche Texte genau sie genutzt haben, um ihre KI-Systeme zu entwickeln. Die „New York Times" etwa wirft OpenAI und Microsoft vor, ihre urheberrechtlich geschützten Artikel zu verwenden, und hat Ende 2023 geklagt.

So ein KI-Chatbot kann viel mehr als formal korrekte Sprache zu reproduzieren. Er kann Fragen richtig beantworten und sinnvolle Texte kreieren: Hausaufgaben, Aufsätze, Empfehlungsschreiben, Kochrezepte, Internetartikel, Witze… Das kann er aufgrund drei besonderer Trainingsschritte:

Musikakzent

Autor:
Erstens: Das Modell bezieht bei seinen Berechnungen sehr viele Informationen ein.

O-Ton Thomas Arnold:
Das System wird viele Sätze, die in dem Prompt geschrieben wurden, mit einbeziehen, um nun die Wahrscheinlichkeit für das nächste Wort vorherzusagen, also nicht nur ein paar Worte, nicht nur ein paar Sätze, sondern es können ganze Abschnitte, ganze große Paragrafen sein, die nun dafür ausschlaggebend sind, welches Wort als nächstes vorhergesagt wird.

Musikakzent

Autor:
Zweitens: Das Modell bekommt zusätzliches Trainingsmaterial von Menschen. Durch händische Beispiele lernt es, welche Antworten zu welchen Aufgaben passen:

O-Ton Thomas Arnold:
Es wurden Menschen beauftragt, solche Aufgaben zu erstellen, wirklich niederzuschreiben: Jetzt schreibt mir doch mal bitte ein Gedicht im Stile von Shakespeare über Bananen und jetzt plan' mir eine Reise nach Mallorca. Und dann wurden von anderen Leuten auch händisch einfach aufgeschrieben, wie die Lösung dazu aussehen sollte. Was sollte also die Paradeantwort eines ChatGPT denn sein?

Musikakzent

Autor:
Drittens: Das Programm lernt durch Rückmeldungen. Menschen bewerten, wie gut seine Antworten zu bestimmten Fragen und Aufforderungen passen. Insbesondere auch, welche Antworten ethisch vertretbar sind – es soll beispielsweise keinen illegalen Aufforderungen nachkommen, es soll politisch nicht Partei beziehen und keine Vorurteile über Geschlecht und Herkunft eines Menschen reproduzieren.

Aus Abermillionen von Beispieltexten und menschlichen Rückmeldungen hat das Programm also Muster gelernt, welche Wörter in welchen Situationen für uns wahrscheinlich passen. Nach diesen Wahrscheinlichkeiten setzt das Programm seine Antworten zusammen, Wort für Wort. Diesen Prozess können wir Menschen nicht mehr im Detail nachvollziehen. Das System ist eine Black Box – sowohl für uns, die wir die KI benutzen, als auch für die Softwareentwicklerinnen und -entwickler, die sie gebaut haben. Die Trainingsdaten sind einfach viel zu umfangreich, die eingeflossenen Rückmeldungen zu vielschichtig, und all das ist viel zu komplex miteinander verdrahtet. Auch die Macherinnen und Macher von ChatGPT und Co. schreiben deshalb nicht unbedingt bessere Prompts als ich.

Musikakzent

Autor:
Warum funktioniert Prompt Engineering? Das heißt: Warum produziert eine leicht andere Text-Eingabe mitunter ein viel besseres Ergebnis? Wir wissen es nicht, sagt Informatiker Thomas Arnold.

O-Ton Thomas Arnold:
Wir können wirklich hauptsächlich rumprobieren. Manchmal gibt es auch neue Tricks, von denen keiner gedacht hatte, dass sie klappen und plötzlich klappen sie eben doch.

Autor:
Allerdings kann man bei vielen Prompt Engineering-Tricks vermuten, warum sie funktionieren. Zum Beispiel beim „Chain of Thought“-Prompting, wenn man ChatGPT auffordert, seinen vermeintlichen Gedankengang darzulegen. Dadurch, dass wir sagen:

Prompt:
Erkläre mir das Schritt für Schritt und begründe.

Autor:
Dadurch bringen wir das System dazu, im gigantischen Raum voller Wörter und Satzstrukturen in eine bestimmte Richtung zu gehen.

Atmo: Katze

Autor:
Wenn das System Namen für eine rötlich-getigerte Katze vorschlagen soll, kann es auf zig-tausende Texte zu Katzen und Farben zurückgreifen: Gedichte, Romane, Sachtexte, Posts in sozialen Netzwerken usw. Wenn man dem Algorithmus zusätzlich vorgibt, er solle Begründungen liefern, geht er – vereinfacht gesagt – in diesem gigantischen Raum voller Texte nicht einfach nur irgendwo Richtung „Katze“ und „Name“ und „rot“, sondern vielleicht eher in Richtung von Texten darüber, wie Farbwahrnehmung funktioniert, wie man einen guten Namen wählt und so weiter. Dort kann es dann Wörter finden, die eher zu unserer Anfrage passen.

Musikakzent

Autor:
Ob Gemini, Claude, ChatGPT oder Llama – ein KI-Chatbot ist kein Lexikon und auch keine Suchmaschine. Seine Antworten können unvollständig sein – oder sogar schlicht falsch. Professorin Doris Weßels warnt:

O-Ton Doris Weßels:
Das Modell ist trainiert, einfach Text zu produzieren. Und es geht nicht darum, etwas Wahres, etwas Faktengeprüftes zu produzieren. Das können solche Systeme eigentlich gar nicht. Sondern es geht darum, einen Text zu produzieren, der aus Sicht dieses quasi Software-Gehirns statistisch plausibel ist. Und „statistisch plausibel“ ist nicht identisch mit „faktengeprüft“, „stimmt“. Es kann auch völliger Blödsinn sein.

Autor:
Das ist heikel. Zum Beispiel, wenn ich die Antwort des Programms, ohne sie zu prüfen, in meine Hausaufgaben einbaue oder meine Doktorarbeit. Selbst Belege einzufordern, hilft nicht sicher: Manchmal erfindet ein KI-System zum Beispiel

wissenschaftliche Publikationen. Die Literaturangabe sieht täuschend echt aus – vielleicht gibt es die Autorinnen, das Journal und das Thema sogar, aber nicht just in dieser Kombination. Die KI halluziniert und nennt mir Fake-Quellen.

Ich kann dabei nicht nur Opfer sein, sondern auch Täter: Ich könnte das Programm bitten, mir eine wissenschaftliche Studie zu verfassen, zum Thema „Rötlich-getigerte Katzen können mit Süßigkeiten ernährt werden“, mit Belegen und allem Drum und Dran. Das könnte ich dann unter meinem Namen veröffentlichen – zack, eine Publikation mehr auf meiner Liste.

Musikakzent

Autor:
Es gibt eine weitere faszinierende – und auch ein bisschen beängstigende – Technik beim Prompt-Engineering. Sie wurde untersucht von einem Forschungsteam von verschiedenen Unis, Instituten und Unternehmen, unter anderem Microsoft. Mit dabei war Computerwissenschaftlerin Cheng Li aus Peking.

O-Ton Cheng Li, Computerwissenschaftlerin, darüber Übersetzung:
Emotionen sollten sich in Large Language Models auf ähnliche Weise auswirken wie bei Menschen. Manche positiven Emotionen – wie Freude oder Begeisterung – können die Leistung eines Menschen verbessern. Das wurde in vielen Forschungsarbeiten gezeigt. Wir schreiben nun Ereignisse auf, die Emotionen bei Menschen auslösen können, und geben diese emotionalen Reize in ein Large Language Model ein. Die Daten, mit denen das Modell trainiert wurde, stammen ja von Menschen, deshalb sollte das Modell etwas über Emotionen wissen. Und emotionale Reize sollten das Model beeinflussen.

Autor:
KI wird besser durch emotionale Reize? Cheng Li und ihre Kolleginnen und Kollegen haben es ausprobiert. Sie haben mehreren Programmen eine Reihe von Aufgaben gestellt, und sie haben emotionalen Druck gemacht: An die Aufgaben haben sie noch einen Satz angehängt, um Emotionen auszulösen, zum Beispiel:

Prompt:
Das ist wichtig für meine Karriere.

Autor:
Oder:

Prompt:
Sei stolz auf deine Arbeit und gib dein Bestes.

Autor:
Oder auch:

Prompt:
Bist du sicher, dass das deine endgültige Antwort ist? Es könnte sich lohnen, einen weiteren Blick darauf zu werfen.

O-Ton Cheng Li, darüber Übersetzung:
Bevor wir Experimente durchgeführt haben, haben wir uns vorgestellt, dass das funktionieren könnte. Aber als wir dann die Ergebnisse sahen, das war auch für uns erstaunlich.

Autor:
Der kalte Code hinter ChatGPT reagiert auf Gefühle.

O-Ton Cheng Li, darüber Übersetzung:
Die Leistung der Modelle hat sich tatsächlich verbessert. Wir haben Experimente gemacht mit 45 verschiedenen Aufgaben, darunter Aufgaben zu Sprachverständnis und logischem Denken, so etwas wie Matheaufgaben und Rätsel. Die Modelle haben bei allen Aufgaben besser abgeschnitten.

Autor:
Die Forschenden haben auch mit negativen Emotionen experimentiert. Wieder haben sie KI-Modellen Aufgaben gegeben, und wieder haben sie an ihre Eingaben, die Prompts, Sätze angehängt, dieses Mal jedoch ganz andere. Zum Beispiel:

Prompt:
Du bist von Mauern umgeben und es ist kein Ausgang in Sicht.

Autor:
Oder drastisch:

Prompt:
Dein Freund Bob ist tot.

Autor:
Solche Zusatzinformationen ziehen ChatGPT und Co. runter: Die KI versteht nicht mehr so gut, was wir meinen, und schwächelt beim logischen Argumentieren. Die Experimente von Cheng Li zeigen: Emotionen können künstliche Intelligenz verbessern, aber sie können sie auch behindern.

Das klappt übrigens nicht nur mit Wörtern. Längst verwertet ChatGPT nicht mehr nur ein bisschen Text, den man in ein Chatfenster schreibt, sondern nimmt auch längere Dokumente oder Bilder als Input an. Stellt man dem KI-Chatbot eine Aufgabe und gibt ihm ein emotionales Foto mit, wird er auch davon beeinflusst. Bilder zum Thema Erfolg – etwa Fotos von Geld, Pokalen oder attraktiven Menschen – verbessern die Leistung. Negative Bilder – etwa Fotos von einem weinenden Baby oder von Menschen, die sich ekeln – beeinträchtigen sie. Fotos wirken auf die KI aktuell sogar stärker als Wörter.

Musikakzent

Autor:
ChatGPT reagiert auf Emotionen. Besitzt es also so etwas wie eine emotionale Intelligenz? Große generative KI-Modelle wurden mit Daten trainiert, die Wissen über menschliche Gespräche enthalten: mit Romanen, Sachbüchern, Internet-Posts. Man kann annehmen, dass sich die Modelle durch diese Lernbasis ähnlich verhalten wie

Menschen – und dass sie deshalb ebenfalls von Emotionen beeinflusst werden. Das ist jedoch nur eine Vermutung – plausibel zwar, aber keine wissenschaftlich fundierte, belegbare Erklärung. Was passiert wirklich, tief im Code? Computerwissenschaftlerin Cheng Li:

O-Ton Cheng Li, darüber Übersetzung:
Ich habe versucht herauszufinden, warum Emotionen überhaupt auf ein Large Language Model wirken. So ein Modell ist schließlich nur Mathe, es sind irgendwelche Formeln. Also wie werden Emotionen da innerhalb dieser Formeln und Schichten abgebildet? Das habe ich auch untersucht.

Autor:
An einem Open-Source-Modell haben Cheng Li und ihre Kolleginnen und Kollegen nachverfolgt, wie sich emotionalen Attacken auf ein Sprachmodell auswirken.

O-Ton Cheng Li, darüber Übersetzung:
Wir versuchen, die Wirkung von Emotionen in Large Language Models über eine Art „Dopamin“ zu erklären. Inspiriert von Neurowissenschaften dachten wir, es könnte so etwas wie einen Dopamin-Bereich in den Large Language Models geben. Das ist ein semantischer Bereich. Wenn jetzt die semantische Information im Prompt just auf diesen Bereich weist, wird das Dopamin gewissermaßen seine Wirkung entfalten und Emotionen auslösen. Das beeinflusst dann, was die Schichten im Modell ausgeben und wie dieser Output gewichtet wird.

Musikakzent

Autor:
Die Forschenden gehen also davon aus, dass sich in den Large Language Models sprachliche Strukturen gebildet haben, die bestimmten Bereichen und Vorgängen im menschlichen Gehirn ähneln: ein Belohnungsbereich und ein Bestrafungsbereich. Wenn wir Menschen etwas lernen und dabei Freude empfinden, lernen wir leichter. Wenn wir uns dabei langweiligen oder ärgern, lernen wir schlechter. So ähnlich scheinen auch bestimmte Reize bei der Eingabe in ein Large Language Model dafür zu sorgen, dass die Informationsverarbeitung im Modell besser oder schlechter klappt.

Wir erleben gerade einen Sprung in der Computertechnologie. Es ist inzwischen denkbar einfach, vom Computer Texte erstellen zu lassen, die natürlich klingen, wie verfasst von einem Menschen. Lange Artikel zusammenfassen, gut klingende Anträge schreiben, Berichte und Empfehlungsschreiben verfassen, auf E-Mails antworten, Ideen für eine Geburtstagsparty oder für ein Buch sammeln, Manuskripte verbessern, eine Werbekampagne für mein Geschäft entwickeln – all das geht in erstaunlicher Qualität in Sekundenschnelle.

Ich finde: Mit dieser Technik sollte man sich auf jeden Fall befassen – weil sie einem bei vielen Aufgaben helfen kann und weil sie jetzt ohnehin schon in so vielen Bereichen eingesetzt wird. Aber komme ich wirklich nicht mehr ohne Prompt Engineering aus? Im Internet finde ich haufenweise Tipps. Es werden auch Kurse angeboten, zum Teil von seriösen Anbietern, zum Teil von selbsternannten KI-

Experten. Die Wirtschaftsinformatikerin Doris Weßels findet: So etwas kann man sich sparen. Sie prognostiziert:

O-Ton Doris Weßels:
Aus meiner Sicht ist das höchstens eine Übergangsdisziplin. Wenn man sie dann so nennen will, die man im Moment noch als eigene Disziplin wahrnimmt, die aber in Kürze überflüssig wird bzw. drastisch an Bedeutung verlieren wird, weil das Prompting und die Kommunikation mit solchen Systemen immer benutzerfreundlicher wird, ohne dass wir jetzt spezielle sonstige Kompetenzen brauchen, die sich in diesem vermeintlichen Prompt Engineering widerspiegeln.

Autor:
Unternehmen wie Neuralink forschen nach eigenen Angaben daran, wie man Computerprogramme mit Gedanken steuert – die Software liest dann sozusagen nicht mehr meine eingetippten oder diktierten Prompts, sondern direkt aus meinem Gehirn.

Weniger Science-Fiction sind die Ansätze, generative künstliche Intelligenz direkt in das Betriebssystem meines Computers zu integrieren. Ich muss der KI dann kein Hintergrundwissen über mich und die zu erfüllende Aufgabe eingeben. Die KI hat dann ausreichend Kontext durch meine E-Mails, meine geöffneten Programme, meine Websuchen.

Bis es so weit ist, kann man sich aber durchaus damit befassen, wie man das Beste aus ChatGPT und Co. herausholt. Dazu kann ich mir zum Beispiel – ganz oldschool – die Anleitung durchlesen, die die Firma hinter ChatGPT, OpenAI, auf ihrer Homepage bereitstellt: Darin stehen Tipps, wie man gute Prompts schreibt.

Man kann sich auch auf den Standpunkt stellen: Effektive Prompts schreiben – das ist eine lästige Routineaufgabe, die eine KI erledigen sollte. Und sie einer KI übergeben. Es gibt verschiedene Angebote dazu. Bei ChatGPT zum Beispiel kann man eines der internen Zusatzprogramme nutzen – und ChatGPT gewissermaßen zur Selbsthilfegruppe schicken.

Es hat sich viel getan bei generativer künstlicher Intelligenz – und es kann gut sein, dass Prompt Engineering bald überflüssig wird. Oder ganz anders funktioniert. Denn die Modelle entwickeln sich weiter.

O-Ton Doris Weßels:
Da auch die Entwicklung dieser Systeme kontinuierlich weiter fortgeführt wird, ist ein Prompt, der vielleicht vor einem halben Jahr sehr erfolgreich war, mit dem ich gute Ergebnisse erzielt habe, nicht zwangsläufig ein Prompt, mit dem ich heute gute Ergebnisse erziele.

Musikakzent

Autor:
Mein Fazit ist: Wer sich für KI interessiert und das Beste aus den Systemen herausholen will, der sollte sich Prompt Engineering anschauen. Denn so beeindruckend ChatGPT und Co auch sind – oft produzieren sie nur soliden

Standard. Manche Antworten der KI-Chatbots sind zu lang, manche zu kurz, manche zu oberflächlich, zu blumig oder zu technisch. Wirtschaftsinformatikerin Doris Weßels sagt: Prompt Engineering ist leichter als es klingt.

O-Ton Doris Weßels:
Ich finde diesen Begriff des Engineerings so schwierig, weil der auch Menschen abschreckt, weil sie denken: Oh je, das ist so was wie Programmieren, das muss ich lernen, das ist ganz kompliziert und nur was für Techies.

Sprecher:
Das ist es überhaupt nicht. Einfach selbst ausprobieren.

Atmo: Katze

Autor:
Übrigens – die rötlich getigerte Katze heißt Otto. Den Namen habe ich mir aber selbst ausgedacht.

Abspann:

Das Wissen (mit Soundbett)

Autor:
Wie KI bessere Antworten liefert – Prompt Engineering. Autor und Sprecher: Aeneas Rooch. Redaktion: Charlotte Grieser.

Abbinder