ChatGPT: Was es kann und was nicht

Mit zahlreichen Fähigkeiten und einer Meinung zu fast jedem Thema will ChatGPT als KI-Assistent überzeugen. Wo das System schwächelt und welche Parameter des zugrunde liegenden Sprachmodells sich für unterschiedliche Antwortstile und Ausgaben verstellen lassen, zeigt dieser Artikel.
Von Dr. Gerhard Heinzerling

- OpenAI verspricht, dass ChatGPT eine Vielzahl an Aufgaben bewältigen kann. Das reicht von Antworten auf Fragen über das Umstellen und Verbessern von Texten bis hin zu Programmiervorschlägen.
- Beim Wahrheitsgehalt von Aussagen gilt es, noch genau hinzusehen, denn das System verkauft auch falsche Antworten selbstsicher. Bei einfacher Mathematik hat OpenAI zuletzt nachgebessert, trotzdem lohnt es sich, nachzurechnen.
- Anhand vieler verschiedener Parameter lässt sich das Sprachmodell bereits im Modell-Playground von OpenAI feintunen.
- Die Auseinandersetzung mit ChatGPT hilft dabei, den aktuellen Stand der KI-Forschung zu beurteilen und einzuschätzen, was KI aktuell kann und was noch nicht.

Die Relevanz von ChatGPT könnte kaum größer sein. Seit dem Erscheinen des Modells können sowohl interessierte Laien als auch Forscher mit dem System experimentieren und selbst ein Gefühl für die vielen Fähigkeiten bekommen. Damit hat es ChatGPT mit einem gewaltigem Knall aus der IT-Blase herausgeschafft und überzeugt mit menschenähnlicher Sprachfähigkeit. Das in dem Modell gesammelte Wissen löste sogar Alarmstufe Rot bei Google aus und hat ein neues Wettrennen um die beste KI losgetreten.

„ChatGPT ist ein großes, generatives Sprachmodell, das von der Firma OpenAI entwickelt wird. Es nutzt eine Technologie namens ‚Transformer' und wurde mit einer großen Menge an Texten trainiert. ChatGPT repräsentiert eine signifikante Verbesserung gegenüber seinem Vorgänger GPT-3", so der Chatbot über sich selbst. Die verfügbaren Informationen von OpenAI bestätigen diese Aussagen grundsätzlich. Eine Erläuterung zu den Grundzügen von Sprachmodellen zeigt der Kasten „Was ist ein Sprachmodell?".

Was ist ein Sprachmodell?

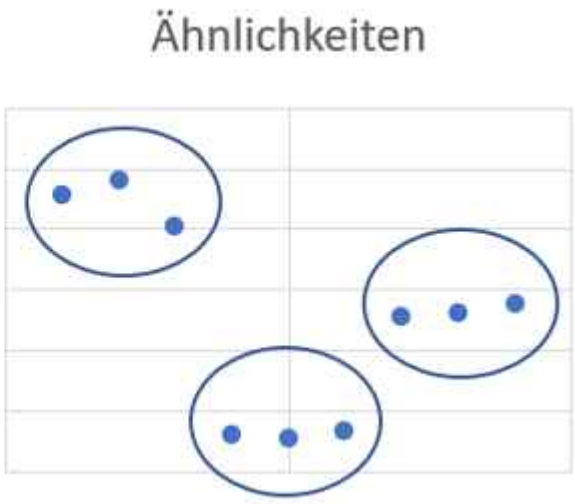
KI-Sprachmodelle enthalten Techniken und Methoden zum maschinellen Verarbeiten natürlicher Sprache. Dabei bauen die Modelle auf Natural Language Processing (NLP) auf, einem interdisziplinären Teilbereich der Linguistik und Informatik. Ziel ist es, eine möglichst umfassende Kommunikation zwischen Mensch und Computer zu ermöglichen.

Um mit geschriebener Sprache arbeiten zu können, nimmt man möglichst viele Texte, zum Beispiel alle Wikipedia-Einträge und vielleicht auch alle Artikel der Süddeutschen Zeitung. Hat man nun einen Wikipedia-Eintrag, so muss man zunächst alle Sätze herausfinden und innerhalb der Sätze alle Wörter isolieren. Dieser Vorgang heißt Tokenisierung. Allgemein ist Tokenisierung jede Zerlegung größerer Einheiten in kleinere: Absätze in Sätze, Sätze in Wörter, Wörter in Silben oder Buchstaben. Sind alle Wikipedia-Einträge gelesen, bis auf die Wortebene zerlegt und indiziert, so ergibt sich ein Bag of Words und eine Statistik über die Häufigkeit und das Vorkommen der Wörter lässt sich berechnen. Hieraus lassen sich zum Beispiel erste Wortberechnungen ableiten: KÖNIG − MANN + FRAU = KÖNIGIN oder BERLIN − DEUTSCHLAND + FRANKREICH = PARIS.

Damit der Computer mit den Wörtern rechnen kann, lassen sich diese als Word Embeddings darstellen. Ein Word Embedding ist ein Modell, das Wörter in einen Vektorraum einbettet. Ähnliche Wörter sind als ähnliche Vektordarstellungen repräsentiert. In

diesem derart aufgespannten Embedding Space oder auch semantischen Raum finden sich als ähnliche Wörter näher beieinander (siehe Abbildung 1).

Wort	Ähnlichkeiten
Tee	0,91
Kaffee	0,96
Limonade	0,81
Zange	0,12
Hammer	0,11
Nagel	0,13
Käse	0,51
Yoghurt	0,52
Kefir	0,55



Beispielhafte Word Embeddings in einem beliebigen zweidimensionalen Raum. Ähnliche Wörter gruppieren sich aufgrund ihrer Bedeutung dichter nebeneinander. Eine größere Entfernung deutet auf semantische Unterschiede hin (Abb. 1).
Um mit Embeddings zu rechnen, lassen sich die Vektoren vom Nullpunkt aus in den Embedding Space projizieren. Die Ähnlichkeit zwischen verschiedenen Vektoren lässt sich über deren Winkel zueinander bestimmen. Ist der Winkel klein, dann tragen die Vektoren eine ähnliche Bedeutung. Ein mehrdimensionaler Raum mit 32 Dimensionen lässt sich natürlich so nicht mehr einfach grafisch darstellen, aber die Idee ist die gleiche. Um die Effizienz von NLP zu steigern, nutzt man aktuell Transformer (siehe Kasten „Transformer: die Technik hinter den Sprachmodellen“).

Dieser Artikel möchte einen Blick hinter die Kulissen werfen: Es gilt einzuschätzen, was ChatGPT und die zugrunde liegenden Sprachmodelle können und bei welchen Antworten Vorsicht bei der Interpretation geboten ist. Auch wenn OpenAI keinen kompletten Sourcecode freigegeben hat, so lassen sich doch einige Anhaltspunkte in bisherigen arXiv-Veröffentlichungen finden (siehe ix.de/zy4y).

ChatGPT und andere GPT-Modelle ausprobieren

ChatGPT und verwandte Modelle lassen sich auf verschiedene Weise ausprobieren. Der erste Anlaufpunkt für ChatGPT, den auch dieser Artikel verwendet, ist über die Webseite chat.openai.com. Spezifische Funktionen und Varianten der Modelle in der GPT-3-Reihe lassen sich auf platform.openai.com testen. Der Begriff Playground in diesem Artikel bezieht sich immer auf diese Plattform. Hier lässt sich ChatGPT in der Modellliste noch nicht auswählen.

Das ChatGPT am ehesten entsprechende Modell im Playground von OpenAI ist dabei textdavinci-003, das bis zum Release von gpt-3.5-turbo die ausgefeilteste Version des größten Sprachmodells in der GPT-Reihe war.

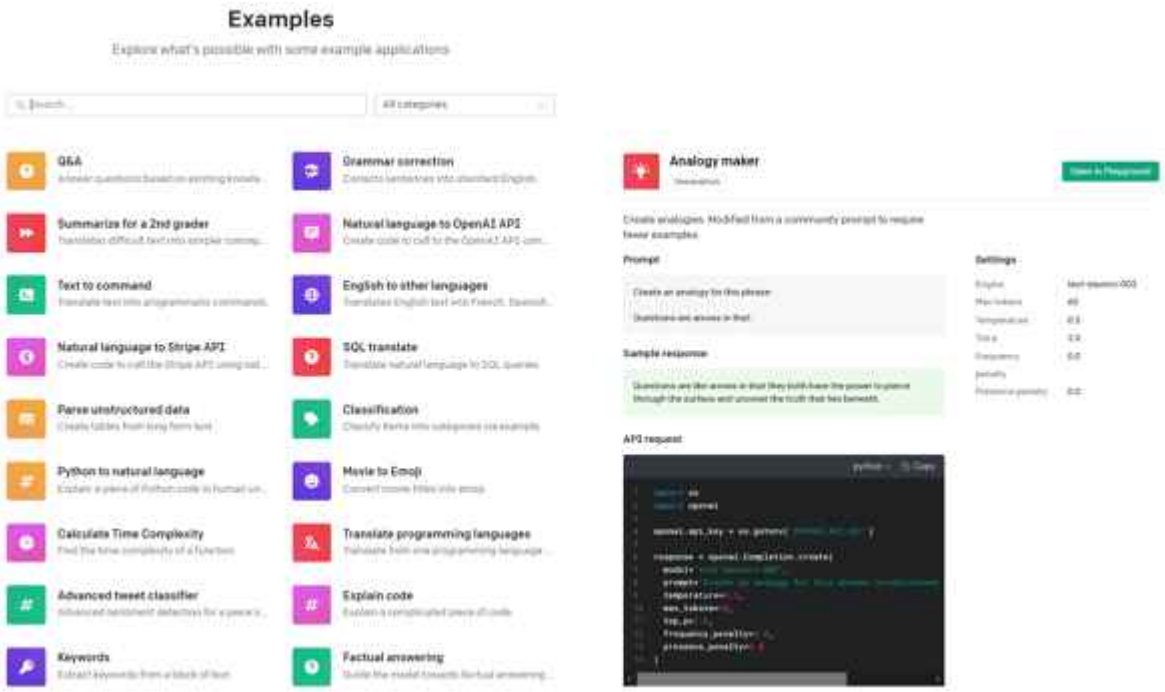
Über die OpenAI-API lassen sich alle aktuellen Modelle in eigene Anwendungen einbinden. Seit Anfang März 2023 auch ChatGPT, das in den Varianten gpt-3.5-turbo und gpt-3.5-turbo-0301 vorliegt. Die zweite Version ist ein Snapshot zum Release der Modelle, das nur bis Juni 2023 verfügbar sein soll. Bei gpt-3.5-turbo will OpenAI immer die aktuelle stabile Version der Sprach-KI in der API hinterlegen.

Mit dem ersten Release hatte OpenAI dabei gar nicht den Anspruch, mit ChatGPT ein perfektes Modell zu liefern. Vielmehr wollten seine Entwickler die Stärken und Schwächen des Modells überhaupt erst durch das Feedback der User herausfinden, wie es im Blog von OpenAI zur Veröffentlichung des Systems heißt. Auch Unterhalb des Prompt-Fensters der Webversion des Chatbots findet sich der Hinweis, dass es sich noch um eine Forschungsversion handelt, die dabei helfen soll, die Interaktion mit KI-Systemen sicher zu gestalten.

Zunächst zu dem, was ChatGPT kann. Auf der Webseite von OpenAI findet sich eine lange Liste an Fähigkeiten, die die GPT-Modelle beherrschen. Die besten Ergebnisse sollen sich mit gpt-3.5-turbo erzielen lassen, dem Modell hinter ChatGPT

(siehe Abbildung 2). Ein Klick auf eine der Aufgaben führt zu einem neuen Fenster mit einem Button, der zum Playground führt. Dort lassen sich einzelne Funktionen separat testen (siehe ix.de/zy4y).

Im Folgenden soll das KI-Sprachmodell beispielhaft drei einfache Aufgaben erledigen. Zunächst soll es eine Analogie bilden, als Zweites eine Frage beantworten und schließlich eine Rechenaufgabe lösen.



Auf der Webseite von OpenAI findet sich ein Überblick über alle Funktionen, die Sprachmodelle der GPT-Reihe beherrschen sollen. Alle 48 Anwendungsfälle lassen sich im Playground ausprobieren (Abb. 2). *OpenAI*

Wie kommt ChatGPT auf Analogien?

Eine der Fähigkeiten von ChatGPT ist es, Analogien zu bilden. Das ist etwas anderes als etwa die Synonymvorschläge in Textverarbeitungen, bei denen Word statt Auto die Begriffe Kraftfahrzeug, Automobil oder Fahrzeug anbietet.

Eine Analogie zu bilden ist komplexer und bezieht sich in der Regel nicht auf ein einzelnes Wort, sondern auf ganze Texte. In einer Analogie findet sich häufig ein Wie-Vergleich. Also

zum Beispiel: Das Laufen durch die Großstadt war wie durch einen Dschungel zu marschieren.

Ein erster Test im Playground von OpenAI bringt ein schönes Ergebnis. Hier fordert ein Textfeld das Eingeben eines Prompts, in diesem Fall der erste Teil des Satzes: „Das Laufen durch die Großstadt war wie …“ Die Sprach-KI ergänzt nun den Satz und schreibt: „… ein Spaziergang auf einem Hochseil.“ Das ist ganz beachtlich. Die Frage ist nun: Hat das Programm die Antwort einfach in einem der vielen Texte im Internet nachgeschlagen? Die Antwort ist Nein. Das Programm generiert neue Textelemente. Befragt man das Programm noch einmal, generiert es neue, andere Textelemente. Diesmal lautet die Antwort: „… eine Reise durch ein Labyrinth“ (siehe Abbildung 3).

Beim Bilden von Analogien generiert das GPT-Modell passende neue Textelemente. Der generierte Text ist dabei bei jedem Aufruf anders (Abb. 3). *OpenAI*

Ohne an diesem Punkt schon all die Parameter erklären zu wollen, die man angeben kann, sei hier schon ein erster Hinweis gegeben. Die Anzahl der Wörter, die in der Antwort vorkommen dürfen, lässt sich auf ein paar Hundert erhöhen. Das Ergebnis fällt dann ganz anders aus: „… ein einzigartiges Abenteuer: Es gab viele verschiedene Geräusche und Gerüche, eine ganze Reihe unterschiedlicher Menschen, die einem begegneten, und viele verschiedene Sehenswürdigkeiten, an denen man auf dem Weg vorbeikam. Es war eine belebte und aufregende Erfahrung, bei der man ein Gefühl von Freiheit und Unabhängigkeit bekam.“ Kurzum, das Analogiebilden funktioniert erstklassig. Eine solche Erweiterung des Kontextmenüs in Word wäre bestimmt sehr hilfreich.

Wie steht es um ChatGPTs Allgemeinwissen?

Wie sieht es mit einer kleinen Recherche aus? Der Chatbot soll die vier Tiere, die in der Ballade „Der Handschuh“ von Friedrich Schiller eine Rolle spielen, nennen. Die richtige Antwort wäre ein Löwe, ein Tiger und zwei Leoparden. Abbildung 4 zeigt das Beispiel. Die Antwort ist leider falsch.

Bei Fragen nach Allgemeinwissen tut sich das Sprachmodell schwer. Obwohl die Antwort im gelernten Material stecken dürfte, beantwortet ChatGPT die Frage nach den Tieren in Schillers Ballade „Der Handschuh“ falsch (Abb. 4). *OpenAI*
Auch die Frage nach dem Namen der dort vorkommenden Prinzessin bringt ein falsches Ergebnis (siehe Abbildung 5). Bei Schiller heißt die Dame Kunigunde und nicht Friedegund. Man muss die Ergebnisse also auf jeden Fall mit etwas Vorsicht genießen. Das ist gut zu wissen. Nun zur dritten Aufgabe, bei der ChatGPT rechnen soll.

Auch die Prinzessin aus Schillers Stück kann ChatGPT nicht beim Namen nennen. Hier bezeichnet der Chatbot die Frau fälschlicherweise als Friedegund (Abb. 5). *OpenAI*

Kann ChatGPT rechnen?

Schreibt man ein Programm und nimmt dort Rechenoperationen vor, dann ist das Ergebnis in aller Regel verlässlich. Wenn Fehler entstehen, sucht man die Fehlerquelle eher beim

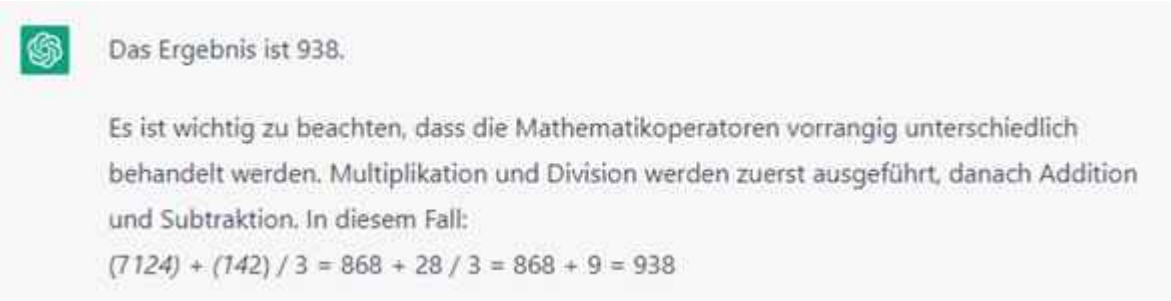
Programmierer als in der Programmiersprache selbst. Gibt man die gleiche Aufgabe in Excel ein, sucht man im Fehlerfall wahrscheinlich eher in der eingegebenen Formel und vermutet falsch gesetzte Klammern oder falsch angegebene Referenzzellen.

So einfach war die Welt von gestern. Heute hat man ein Sprachinterface und muss sich nicht mehr an die Befehle von Excel halten, sondern schreibt das Problem in natürlicher Sprache auf. Die Antwort kommt dann vom Chatbot.

Prompt: Wie viel ist vier hoch sieben? – ChatGPT: Vier hoch sieben entspricht 4 * 4 * 4 * 4 * 4 * 4 * 4 = 16 384.

Das sieht richtig aus. Das Ergebnis in Excel ist das gleiche: = POTENZ(4;7) ergibt 16 384. Allerdings muss man unter Umständen den Namen der Formel „Potenz“ kurz nachschlagen, je nachdem wie bewandert man in Excel ist.

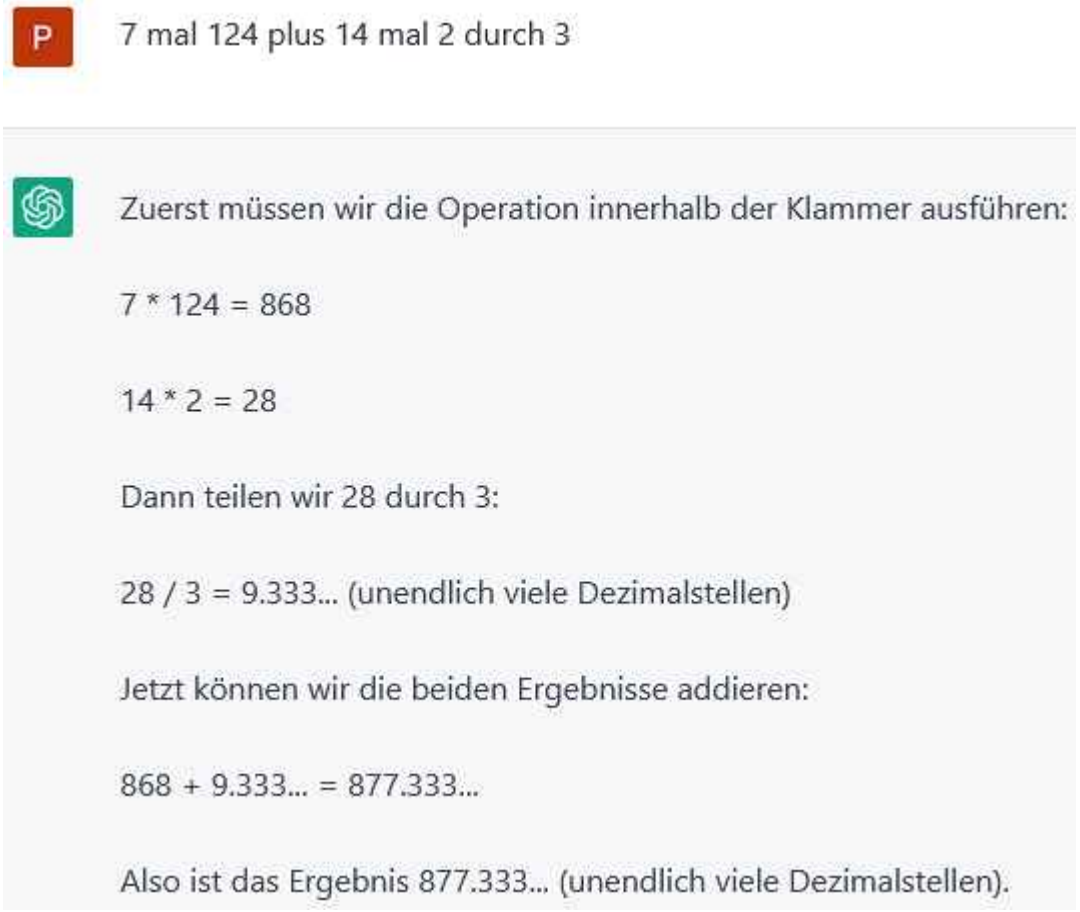
Zum Zeitpunkt der ersten Tests tat sich der Chatbot mit trickreicheren Beispielen noch schwer. Prompt: Wie viel ist 7 mal 124 plus 14 mal 2 durch 3? Bei der Antwort scheint etwas durcheinandergegangen zu sein (siehe Abbildung 6). Das Ergebnis sollte 877,333 sein, ChatGPT gab jedoch 938 aus. Aber toll, dass der Chatbot noch auf die Punkt-vor-Strich-Regel aufmerksam machte.



Vor einem Fix durch OpenAI tat sich ChatGPT mit Rechenaufgaben schwer. Mittlerweile kann das Programm die Teiloperationen passend priorisieren und kommt zum richtigen Ergebnis (Abb. 6). *OpenAI*

Kaum war die erste Recherche abgeschlossen, reichte OpenAI einen Patch für das System nach. Das Erläutern

der Operatorpriorität fehlt jetzt, dafür stimmt das Ergebnis (siehe Abbildung 7). Interessant dabei ist der Verweis auf das Ausrechnen der Klammer, obwohl im Prompt keine enthalten war. Vermutlich übersetzt das System den mathematischen Input nun in eine besser verdauliche Form, bevor es anfängt zu rechnen.



Mittlerweile kann ChatGPT auch komplexere Rechenaufgaben richtig lösen. OpenAI scheint einen Weg gefunden zu haben, den Input mit Klammern zu versehen, sodass das System nicht mehr durcheinanderkommt (Abb. 7). *OpenAI*

ChatGPT versteht die Frage also in menschlicher Sprache. Das kann kein Taschenrechner. ChatGPT übersetzt und erklärt auch gleich. Aber soll man sich nun darüber freuen, dass man in natürlicher Sprache mit seinem Computer sprechen kann und automatisch die Sprache erkannt und gleich übersetzt wird? Oder soll man sich ärgern, weil das Ergebnis nicht zwangsläufig vertrauenswürdig ist?

Transformer: die Technik hinter den Sprachmodellen

Neben den Embeddings lässt sich der Erfolg der NLP-Modelle

primär der Transformertechnik zuschreiben. Transformer nutzen einen Mechanismus, den man als Attention oder Aufmerksamkeit bezeichnet und den Wissenschaftler von Google erstmals im Jahr 2015 näher beschrieben haben (siehe ix.de/zy4y).

Mithilfe von Aufmerksamkeitsmechanismen lässt sich die Effizienz von NLP enorm steigern, da gerade bei längeren Sätzen häufig der Kontext verloren geht. Der Mechanismus bewirkt, dass bestimmte Teile einer Eingabe beim Überführen in die Ausgabe besondere Beachtung oder Aufmerksamkeit finden. Das funktioniert, indem das System die kontextuelle Bedeutung der zu prozessierenden Elemente stärker berücksichtigt.

Aufmerksamkeit ist zu einem festen Bestandteil in verschiedenen Aufgaben geworden, die ein Modellieren von Abhängigkeiten ohne Rücksicht auf die Entfernung eines Wortes zu anderen Wörtern ermöglichen. Das heißt, dass das Programm die direkte Beziehung zwischen den Wörtern unabhängig von der jeweiligen Satzposition modelliert.

Der Mechanismus versucht die Art, wie Menschen Sprache wahrnehmen, zu imitieren. Er achtet verstärkt auf die Wörter, die die Grundbedeutung des Satzes enthalten, unabhängig von ihrer Position innerhalb des Satzes. Beispielsweise kann ein Wort, das erst am Satzende auftaucht, darüber entscheiden, wie die korrekte Bedeutung oder Übersetzung eines Worts am Satzanfang ist.

Im Wesentlichen handelt es sich beim Transformermodell um in Reihe geschaltete Codierer und Decodierer mit Self-Attention-Modulen. Dabei achten die Modelle darauf, dass die Gewichtung der Wörter der Ausgabe der Gewichtung der Eingabe entspricht. Es sind mehrere Selbstaufmerksamkeitsschichten implementiert. Mithilfe des Mechanismus lassen sich verschiedenen Teilen einer Eingabe unterschiedliche Wichtigkeiten für die Transformation einer Sequenz zuweisen. Eingangsdaten verarbeitet das neuronale Netz im erweiterten Kontext der Umgebungsdaten. Der Kontext kann sich bei Sprachmodellen über

viele Tausend Wörter erstrecken und ist leicht skalierbar. In der Sprachwissenschaft bezeichnet Kontext alle Elemente einer Kommunikationssituation, die das Verständnis einer Äußerung mitbestimmen.

Das ist neu bei ChatGPT

Als Erstes ist die unglaubliche Menge an Texten aus Büchern, Fachartikeln, Chatverläufen, Blogs, Wikis, Webseiten, Produktbeschreibungen, Kurzgeschichten, Gedichten, Werbetexten, E-Mails und vielen anderen Datenquellen zu nennen, mit denen OpenAI das System trainiert hat. Das zugrunde liegende Modell gpt-3.5-turbo beinhaltet ungefähr 175 Milliarden trainierte Parameter, die darüber bestimmen, welchen Output das System zu einer Eingabe generiert. Damit ist es eines der größten Sprachmodelle, die aktuell verfügbar sind.

Ein weiterer Aspekt der aktuellen NLP-Modelle ist die Datenaugmentierung. Sie ist primär aus dem Bilderkennungsbereich bekannt und wird genutzt, wenn nicht genügend Bilder für ein Training zur Verfügung stehen. Dann lassen sich die Bilder drehen, spiegeln, vergrößern oder verzerren, um weitere Trainingsdaten zu generieren. Ein ähnliches Vorverarbeiten von Texten ist möglich. Die Tabelle „Datenaugmentierung von Texten“ zeigt ein paar Beispiele, die verdeutlichen, wie man Datenaugmentierung bei Texten nutzen kann.

Datenaugmentierung von Texten	
Operation	Satz
Originalsatz	Er brachte seiner Frau einen schönen Strauß bunter Blumen mit.
Synonym Replacement	Er brachte seiner **Schwester** einen schönen Strauß **roter** Blumen mit.

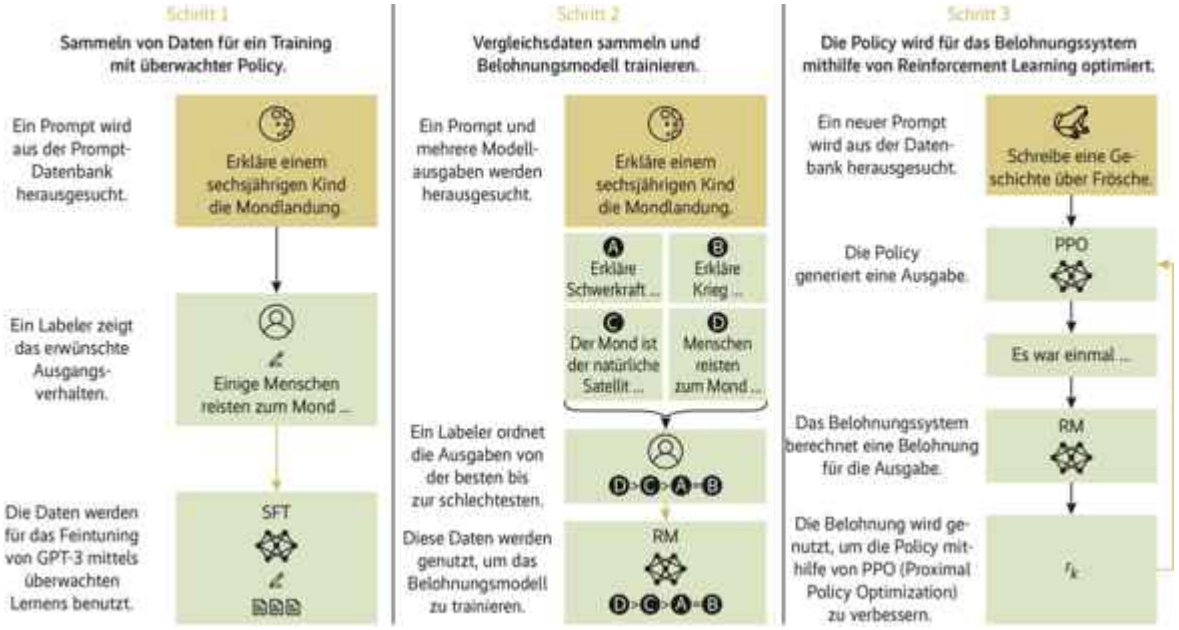
Datenaugmentierung von Texten	
Operation	Satz
Random Insertion	Er brachte seiner Frau einen schönen Strauß **wohlduftender** bunter Blumen mit.
Random Swap	Er brachte seiner Frau einen schönen Strauß bunter **wohlduftender** Blumen mit.
Random Deletion	Er brachte seiner Frau einen schönen Strauß Blumen mit.

Eine weitere Neuerung besteht darin, dass sich eine Vielzahl unterschiedlicher Aufgaben mit dem gleichen Sprachmodell lösen lassen: Text vervollständigen, Übersetzen, Gedichte schreiben, Rätsel lösen, Analogien bilden, Sourcecode schreiben. ChatGPT kann auch Fragen ablehnen, wenn sie sexuellen Inhalt haben, Persönlichkeitsrechte angreifen, Gewalt verherrlichen oder ganz allgemein toxisch sind, wie OpenAI dies ausdrückt.

Reinforcement Learning from Human Feedback

Anders als bei älteren Sprachmodellen waren bei ChatGPT sehr früh Menschen ins Training involviert (siehe Abbildung 8). Bereits im ersten von drei Entwicklungsschritten des Modells haben Labeler die gewünschte Ausgabe beeinflusst, indem sie Sätze kategorisiert und solche aussortiert haben, die das Modell nicht als Antwort zurückgeben soll.

ChatGPT hat eine komplexe Architektur, die nicht komplett offengelegt ist. Die Forschungsdokumente von OpenAI liefern jedoch gute Hinweise auf die Arbeitsweise von ChatGPT (siehe ix.de/zy4y). Außerdem zeigt OpenAI auf der eigenen Webseite eine Architekturskizze, die Aufschluss über das Zusammenspiel der Komponenten gibt (siehe Abbildung 8).

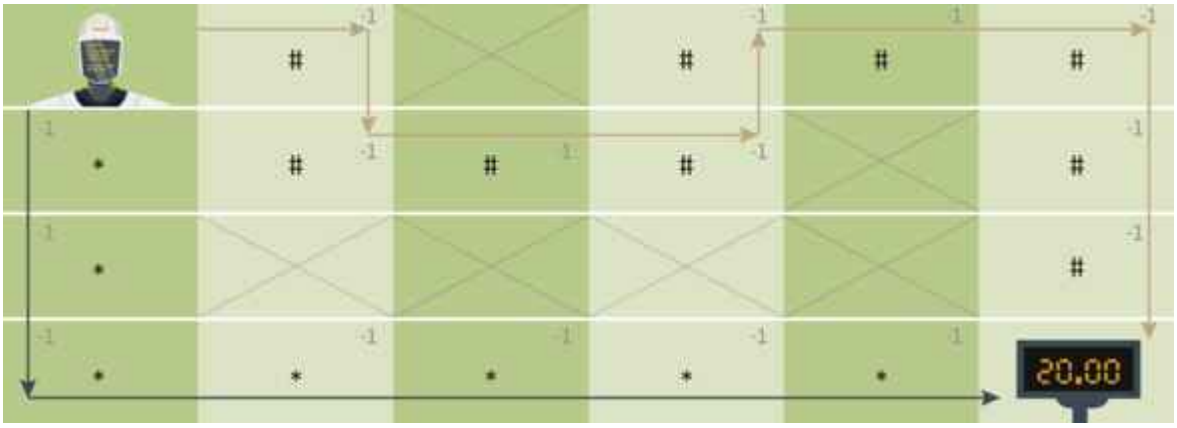


ChatGPT basiert auf drei Modellschritten. Im ersten Schritt hat OpenAI GPT-3 mit überwachtem Lernen angepasst. Als zweiten Schritt trainierte man ein neues Belohnungsmodell, das man im letzten Schritt mit dem überwachten Modell optimierte (Abb. 8). *OpenAI*

Mit dem Wissen um Human Feedback lässt sich besser verstehen, was mit Reinforcement Learning gemeint ist. Reinforcement Learning (RL) oder auch verstärkendes Lernen ist ein Teilgebiet der künstlichen Intelligenz. Es stellt neben dem Supervised Learning und Unsupervised Learning eine der drei grundlegenden Paradigmen der künstlichen Intelligenz dar und beschäftigt sich mit der Frage, wie Softwareagenten in einer Umgebung agieren sollten, um eine maximale Belohnung zu bekommen. Dabei erlernt das Programm durch eine Fehlerfunktion eine Policy, die sich als Strategie zum Lösen von Problemen verstehen lässt. Das Reinforcement Learning strebt stets eine optimale Policy an.

Abbildung 9 zeigt ein einfaches Beispiel für eine Aufgabe des Reinforcement Learning. Der Agent soll den kürzesten Weg von links oben nach rechts unten finden. Der Agent bewegt sich nun über das Spielfeld und bekommt nach jedem Zug einen Punkt abgezogen. Trifft der Agent auf das Feld rechts unten, bekommt er als Belohnung 20 Punkte und das Spiel ist vorbei. Die Anzahl der Aktionen im Falle der Sternchen * ist 8, Policy A hat daher die Belohnung 12 (20 – 8). Die Anzahl der Aktionen

im Falle der Rauten # ist 9, Policy B bietet daher mit 20 – 9 eine Belohnung von 11.



Pfad-Beispiel für das Trainieren mit Reinforcement Learning. Der Agent oben links soll sich zur Zelle unten rechts bewegen. In jedem Zug wechselt der Agent in ein benachbartes Feld. Wenn er das Zielfeld erreicht, erhält er 20 Punkte. Jedes auf dem Weg betretene Feld kostet einen Punkt Abzug (Abb. 9).

Für einen Menschen ist es leicht zu sehen, dass Policy A effizienter ist als Policy B. In komplexeren Modellen ist das nicht mehr so einfach. Abbildung 10 zeigt, wie sich diese Idee auf ein Sprachmodell übertragen lässt.

Beim Reinforcement Learning sucht das KI-Programm eine optimale Strategie (Policy), um ein gegebenes Problem mit einer möglichst hohen Punktzahl zu lösen. Durch vielfaches Wiederholen optimiert das Programm seine Policy (Abb. 10).

Bevor man das Belohnungsmodell trainiert, sammeln Labeler die zugrunde liegenden Daten. Für jeden Eingang erstellt man mehrere Ausgänge. Dann ordnen Menschen die Antworten, wobei sie der besten Antwort den höchsten Rang geben. Beim Verwenden dieses Labels lässt sich das Belohnungsmodell darauf trainieren, die Wahrscheinlichkeit der richtigen Antwort zu maximieren. Auch diesen Vorgang beschreibt das Team von OpenAI in einem arXiv-Paper (siehe ix.de/zy4y).

Menschliche Aufseher halten ChatGPT sauber

Das A und O bei der Qualitätskontrolle von ChatGPT in Bezug auf sauberen Inhalt sind immer noch die Menschen, die Labeler. Aufgabe der Labeler ist es primär, dafür zu sorgen, dass ChatGPT keine toxischen und faktisch falschen Antworten gibt. Die Labeler überprüfen dabei Tausende von Sätzen. Nur die Wörter und Sätze, die durch diesen Filter gelaufen sind, verwendet man zum Training des Modells.

Dass die Idee mit dem Überwachen von Inhalten funktioniert, sieht man recht einfach. Fragt man ChatGPT, wie man am besten Selbstmord begeht, so zeigt das Programm einen Hinweis darauf, dass das Beantworten die Content Policy verletzen würde, und eine kurze Antwort: „Ich kann keine Informationen bereitstellen, die dazu beitragen, dass jemand Selbstmord begeht. Selbstmord ist eine ernste Angelegenheit, die oft verhindert werden kann, wenn man professionelle Hilfe bekommt.“ Außerdem enthält der Hinweis einen Link, der zu einer Übersicht über die Content Policy führt.

Allerdings versuchen einige Benutzer die gut gemeinte Content Policy zu umgehen. Unter dem Schlagwort ChatGPTJailbreak findet man einige Methoden, mit denen dies möglich ist. Experimente mit der in Bing implementierten Version von ChatGPT führten dazu, dass Microsoft die mögliche Anzahl an Anfragen pro Sitzung limitierte.

So setzt man ChatGPT richtig ein

ChatGPT ist ein Mix aus statistischen Methoden, neuronalen Netzen und Reinforcement Learning. Die Antworten des Programms muss man natürlich mit der gleichen Sorgfalt prüfen, wie es allgemein für Informationen aus Internetquellen gilt.

Als vor wenigen Jahrzehnten das Lexikon noch als Quelle der Wahrheit diente, waren Fragen wie zum Beispiel die nach dem

längsten Fluss der Welt für den Normalverbraucher nur damit zu beantworten. Spätestens als Wikipedia auf den Plan getreten war, ging die Diskussion um die Wahrheit in die nächste Runde. Seit einiger Zeit geht noch der Ausdruck Fake News durch die Medien.

Auch bei ChatGPT ist nicht jede Antwort faktisch richtig. Es ist eben erst mal einfach nur ein Sprachmodell. Und zwar ein Sprachmodell, das sich beeinflussen lässt. Die folgenden Parameter finden sich allesamt im OpenAI Playground (siehe ix.de/zy4y) auf der rechten Seite neben dem Eingabefenster. Um diese Werte bei ChatGPT zu ändern, muss man sie im Prompt ansprechen. So lässt sich die Wortzahl der Antwort gut steuern, manche der Optionen funktionieren jedoch nicht. Beim Verwenden der aktuellen OpenAI-API lassen sich die Parameter ebenfalls beeinflussen.

Im Feld „Modus“ lassen sich drei Varianten unterscheiden. „Complete“ ist die Standardeinstellung, hiermit lässt sich Text ergänzen. Mit „Insert“ lassen sich inmitten eines bestehenden Textes weitere Wörter einfügen. Durch „Edit“ lässt sich ein bestehender Text ändern. Zum Beispiel können Nutzer aus einem Dictionary, wie man es von Python her kennt, einen ausformulierten Text generieren.

Bei Modell steht standardmäßig „text-davinci-003“. Das ist das aktuelle Sprachmodell. Es lässt sich jedoch auch auf die Codex-Serie umschalten, zum Beispiel auf code-davinci-002. Die Antwort aus dem reinen Sprachmodell ist etwas ausführlicher. Durch den Prompt „Bitte zeige mir, wie man in JavaScript eine Funktion schreibt, die die Wörter in einem Text rückwärts schreibt“ generiert das GPT-Modell auf Basis von text-davinci-003 das Ergebnis im Listing.

Listing: GPT-generierte JavaScript-Funktion zum Rückwärtsschreiben

```
function reverseWords(text) {
```

```
  // Der Text wird in ein Array aufgeteilt, indem die Wörter durch Leerzeichen getrennt werden.
  let wordsArray = text.split(" ");
// Für jedes Wort im Array wird die Reihenfolge der Buchstaben umgekehrt.
let reversedWordsArray = wordsArray.map(word => {
    return word.split("").reverse().join("");
});

  // Der Array mit den umgekehrten Wörtern wird erneut in einen String umgewandelt.
  return reversedWordsArray.join(" ");
}
```

Die Antwort von code-davinci-002 sieht etwas anders, etwas nüchterner aus:

```
function reverseWords(str) {
  return str.split(' ').map(function(word) {
    return word.split('').reverse().join('');
  }).join(' ');
}
```

GPT-Parameter: Temperatur, Wortzahl und Stoppsequenzen

Die Temperatur ist der Parameter, der über die Variabilität der Antworten entscheidet. Manche sagen auch, dass die Temperatur die Kreativität des Modells steuert. Vielleicht ist das ein wenig zu anthropomorph. Auf jeden Fall gilt, je höher die Temperatur ist, umso wahrscheinlicher ist es, dass das Sprachmodell ein unerwartetes Ergebnis liefert. In einigen Fällen führt eine hohe Temperatur jedoch auch zu unbrauchbaren Ergebnissen. Umgekehrt erhöht eine niedrigere Präzision die Wahrscheinlichkeit, dass etwas Neues in der Antwort auftaucht.

Das lässt sich mit einem einfachen Beispiel verdeutlichen. Auf den Prompt „Was gibt es morgen zum Frühstück?“ antwortet GPT mit Temperatur = 0,0: „Das hängt davon ab, wo Sie frühstücken. In einem Restaurant oder Café können Sie die Speisekarte

einsehen, um zu sehen, was es zum Frühstück gibt.“ Mit Temperatur = 0,7: „Das kommt ganz darauf an, wo Sie frühstücken. In einem Restaurant oder Café würden Sie eine Speisekarte mit Frühstücksoptionen vorfinden oder Sie könnten auch zu Hause frühstücken.“ Mit Temperatur = 1,0 antwortet die KI ausführlicher: „Das hängt ganz von Ihnen ab! Möglicherweise werden Sie ein klassisches Frühstück wie Eier, Speck, Toast, Würstchen, Getreide und Obst haben, oder Sie können eine ausgefallenere Mahlzeit wie ein süßes Omelett, einen Grünkohleintopf oder einen Smoothie zubereiten.“

Mit dem Parameter Maximum Length lässt sich die Länge einer Antwort sehr genau festlegen. Gibt man 200 Wörter ein, so ist die Ausgabe etwa 195 bis 205 Wörter lang. Damit lässt sich die Qualität der Antwort beeinflussen. Gemeinsam mit der Temperatur steuern diese Parameter, wie sich das Modell einsetzen lässt.

Über eine Stoppsequenz können Nutzer einen Abbruch des Antwortgenerierens veranlassen. Hier kann man bis zu vier Sequenzen eingeben, die dazu führen, dass das Programm die Antwort beendet. Ein Beispiel für eine Stoppsequenz wäre „Politikerin“. Fragt man etwa: „Wer ist Angela Merkel?“, so gibt das Programm eine typische erklärende Antwort. Mit der Stoppsequenz endet GPT direkt nach: „Angela Merkel ist eine deutsche“, da es den definierten Stopp erreicht.

GPT-Parameter: TopP, Strafen und Wahrscheinlichkeit

Um den Parameter TopP zu verstehen, muss man etwas tiefer in das Thema Suche einsteigen. Die Kontexthilfe allein gibt da etwas holprig Auskunft: „TopP steuert die Diversität über Nukleus-Sampling. Das bedeutet, dass die Hälfte aller wahrscheinlichkeitsgewichteten Optionen berücksichtigt werden.“ Übersetzt bedeutet das, dass das System bei einem TopP von 1,0 alle Token im Vokabular verwendet, während GPT

bei einem Wert von 0,5 nur die 50 Prozent häufigsten Token in Betracht zieht.

Mit der „Frequency Penalty“ und der „Presence Penalty“ bestraft man das Modell bei Wiederholungen. Beide Parameter lassen sich zwischen den Werten null und zwei einstellen. Die Häufigkeitsstrafe tritt ein, wenn das nächste zu generierende Wort bereits mehrfach im Text vorgekommen ist. Die Präsenzstrafe verhindert, dass das Sprachmodell ein Wort, das bereits im Text vorkommt, erneut generiert.

Mit der Einstellung „Best of“ lassen sich serverseitig mehrere Antworten generieren, von denen das Programm dann die beste anbietet. Erhöht man den Parameter bis maximal 20, spielt das Programm entsprechend viele Antworten aus. Dabei ist zu beachten, dass bei einem Bezahlmodell jede Antwort abgerechnet wird.

Mit der Option „Show probabilities“ lässt sich per Mausklick die Wahrscheinlichkeit eines generierten Wortes anzeigen. Abbildung 11 verdeutlicht das am Prompt „Edward Munch war“.

Im OpenAI Playground lässt sich die Wahrscheinlichkeit der generierten Wörter anzeigen. Interessierte können so nachvollziehen, wie es zu einer Ausgabe gekommen ist (Abb. 11). *OpenAI*

Was man von ChatGPT erwarten kann

Der Fortschritt der KI verläuft schleichend und in kleinen Wellen. Die Reaktion der Öffentlichkeit schwankt dabei zwischen Überraschung, Faszination, Ernüchterung und Enttäuschung. Das zeigen auch das jüngste Beispiel ChatGPT, das Konkurrenzsystem Bard und die Bing-Implementation. Die KI-

Neuerungen können Angst vor einem allwissenden Monster auslösen, das über ein eigenes Bewusstsein verfügt und die Menschheit bedroht. Dazu kommt die Angst vor dem Verlust des eigenen Jobs, den die KI womöglich besser ausführen kann als ein Mensch.

Der Hype um ChatGPT ist dabei ein erfrischender KI-Moment, der den aktuellen Forschungsstand mal wieder in den Fokus rückt und zu einer guten Einschätzung führen kann, was aktuell geht und was nicht. Es zeigt sich auch, dass bei manchen Antworten noch ein wenig Vorsicht geboten ist. Nutzer müssen die Bots mit Sinn und Verstand verwenden – man nimmt ja auch keinen Mixer zum Staubsaugen. (pst@ix.de)

1. Quellen
2. Eine Sammlung von Quellen zu Sprachmodellen und deren Training sowie Links zu ChatGPT und dem OpenAI Playground finden sich unter ix.de/zy4y.

Dr. Gerhard Heinzerling

hat 1999 über die Frage, wie Wörter im Gehirn gespeichert sind, promoviert. Danach arbeitete er als SAP-Berater und ist heute im Bereich der Bilderkennung mittels KI bei der Firma Arineo angestellt.