Domain-Specific Training: Der heilige Gral für ChatGPT & Co.

ChatGPT & Co. können menschliche Sprache ziemlich gut. Sie können zum Teil auch Inhalt – zumindest wenn es sich um allgemeines Wissen handelt. Aber können sie auch komplexes Spezialwissen? Das ist die spannende Frage, die sich in der nächsten Zeit entscheiden wird.

Kann man Large Language Models so trainieren, dass sie auch auf Spezialfragen mit hoher Wahrscheinlichkeit richtige Antworten geben? Symboldarstellung von KI-Chatbots in einem Callcenter. Das Bild wurde mit dem Bildgenerator Leonardo AI erstellt.

Können Large Language Models wie GPT, Googles LaMDA oder die frei verfügbaren GPT-Neo und LLaMA anhand von Spezialwissen einfach trainiert werden und dann korrekte Antworten liefern? Kann ich ein nichtproprietäres Large Language Model downloaden, das Model dann mit einem Gesetzestext oder einer Gebrauchsanleitung, den Informationen meiner Homepage sowie ein paar Hundert E-Mails vom Kundendienst füttern, sodass es auf Anfragen richtige Antworten liefert – und zwar mit einer Trefferquote von mindestens 80 %, während gleichzeitig falsche Antwort nahezu ausgeschlossen werden? Im Fachjargon wird das Domain-Specific Training oder Fine-Tunig genannt. Das ist der heilige Gral. Das ist es, was Wirtschaft und Verwaltung wirklich brauchen, um die Produktivität auf ein völlig neues Niveau zu heben. Auf diese Aufgabe stürzen sich derzeit zahllose Forscher.

Wahrscheinlichkeiten statt Schlussfolgerungen

Werfen wir einen Blick in den Motorraum der Sprachmodelle: Im Endeffekt produziert das Modell, abhängig vom Eingabetext, eine Liste von Wörtern, die dem Eingabetext folgen und nach Wahrscheinlichkeiten geordnet sind. Auf die Frage «Wie funktioniert der Wasserkreislauf?» könnte es etwa die folgende Liste produzieren: «Der»: 4,2 Prozent, «Die»: 2,9 Prozent, «Das»: 2,8 Prozent, «Ein»: 1,9 Prozent und so weiter. Sagen wir, das Modell wählt «Der» aus. Dann erstellt es im nächsten Schritt eine Liste der wahrscheinlichsten Worte für den Eingabetext: «Wie funktioniert der Wasserkreislauf? Der» und wählt wieder ein Wort aus, zum Beispiel «Wasserkreislauf». Dann sind wir schon bei «Wie funktioniert der Wasserkreislauf? Der Wasserkreislauf». Und so hangelt sich das Modell Wort für Wort und Wahrscheinlichkeit für Wahrscheinlichkeit immer weiter. Es geht dabei also um statistische Vorhersagen, nicht um Schlussfolgerungen oder Argumentationen.

GPT wählt nicht immer das Wort mit der höchsten Wahrscheinlichkeit aus. Macher von ChatGPT haben einen Parameter eingebaut, den sie «Temperature» nennen. Er bestimmt, wie oft das Modell ein Wort mit einer niedrigeren Wahrscheinlichkeit auswählen kann. In unserem Beispiel könnte sich das Modell für «Das» oder «Ein» anstelle von «Der» entscheiden und damit weitermachen: «Wie funktioniert der Wasserkreislauf? Das Wasser …» In der Temperature liegt auch einer der Gründe, warum ChatGPT auf die gleiche Frage immer leicht abgewandelte Antworten liefert. Es gibt keine Theorie, warum die Temperature notwendig ist, nur die Beobachtung, dass sie funktioniert. Genau genommen muss das Modell künstlich ein unsichereres Ergebnis wählen, damit die Texte menschengemacht erscheinen. Dieser Umstand könnte – Achtung, jetzt wird es spekulativ – die generelle Fehlerhaftigkeit oder Irrationalität zumindest der menschlichen Sprache enthüllen. Oder auch, dass kreative Intelligenz und Fehlerfreiheit sich generell ausschliessen.

Es funktioniert – aber warum?

Spannend ist auch die Frage, woher die Wahrscheinlichkeiten kommen. Wenn man das gesamte Internet und alle digitalisierten Bücher – zusammen ein paar Hundert Milliarden Wörter – auswertet, kann man zwar herausfinden, mit welcher Wahrscheinlichkeit bestimmte Wörter vorkommen. Man kann aber keine Wahrscheinlichkeiten für eine Kette von Wörtern identifizieren. Dafür gibt es kombinatorisch einfach viel zu viele Möglichkeiten. Wir können diese auch nicht mithilfe von Formeln berechnen. Aber wir können ein künstliches, neuronales Netzwerk bauen, welches für uns die Wahrscheinlichkeiten überzeugend schätzt. Wir verstehen zwar nicht, wie das neuronale Netzwerk Muster erkennen kann oder welche Gesetzmässigkeiten es genau identifiziert hat, wir beobachten aber, dass es funktioniert. Das neuronale Netzwerk ist für uns eine Blackbox. Wir geben etwas hinein, spielen ein bisschen an den Gewichtungen herum, geben Feedback und erhalten eine Ausgabe, die uns verblüfft.

Sprache und Inhalt sind jedoch zwei unterschiedliche Paar Schuhe. Woher kommt nun der Inhalt? Wieso kann ChatGPT inhaltlich präzise und korrekt auf die Frage «Wie funktioniert der Wasserkreislauf?» antworten? Zunächst einmal, weil es genug Texte gesehen hat, die den Wasserkreislauf korrekt beschreiben – und nur sehr wenige, die ihn falsch beschreiben. Zum anderen, weil es darauf und auf den gesamten Trainingstext aufbauend genügend Gesetzmässigkeiten erkannt hat, um entsprechende Listen von Wahrscheinlichkeiten für die jeweilig nächsten Worte zu erstellen, die auch einen inhaltlich korrekten und nicht nur sprachlich präzisen Text ergeben. Damit wird auch klar, warum es auf andere Fragen weniger gut antworten kann: Wenn es um spezielles Wissen geht, von dem es nur wenige Texte gibt oder um Sachverhalte, die umstritten sind und zu denen es inhaltlich unterschiedliche Texte im Internet gibt, schätzt es die Wahrscheinlichkeiten weniger treffsicher.

> **Das künstliche neuronale Netzwerk kann nicht zwischen einer inhaltlich korrekten und einer sprachlich wahrscheinlichen Antwort unterscheiden.**

Klingt gut, ist aber falsch

Damit zurück zur Ausgangsfrage. Was ist, wenn man relativ wenig Text (Trainingsdaten) zur Verfügung hat, in denen dafür kompliziertes Spezialwissen steckt? Jetzt wird es richtig interessant. Können Large Language Models ausschliesslich Muster in der menschlichen Sprache erkennen, wird das Domain-Specific Fine-Tuning mit wenigen Informationen nicht möglich sein, zumindest nicht mit einer ausreichend hohen Genauigkeit und ohne fehlerhafte Antworten. Das Modell wird aus den wenigen Daten zwar Wahrscheinlichkeiten für das nächste Wort berechnen können, die einen sprachlich korrekten Text liefern, der nach Spezialwissen klingt, der aber inhaltlich nicht stimmt. Das liegt etwa daran, dass in Texten mit Spezialwissen, die zentralen Fachbegriffe immer wieder in vielen verschiedenen Kontexten erscheinen. Das künstliche neuronale Netzwerk unterscheidet aber nicht nach Kontext und Inhalt, es produziert die Wahrscheinlichkeiten anhand des erkannten sprachlichen Musters. Daher kann es nicht zwischen einer inhaltlich korrekten Antwort und einer sprachlich wahrscheinlichen Antwort gemäss dem erkannten statistischen Muster unterscheiden. Es wird daher häufig «halluzinieren», wenn inhaltlich korrekt und sprachlich wahrscheinlich nicht deckungsgleich sind: Der Ausgabetext wird richtig klingen, obwohl er es gar nicht ist. Genau darin liegt auch die grosse Gefahr. ChatGPT und ähnliche Anwendungen könnten Antworten hervorbringen, die Laien und Halbwissende fälschlicherweise überzeugen und so für viel Verwirrung sorgen und jede Menge Probleme schaffen, anstatt sie zu lösen.

Schlüssel zum menschlichen Denken?

Neueste Forschungsergebnisse deuten an, dass Domain-Specific Training beziehungsweise Fine-Tuning, basierend auf frei verfügbaren vortrainierten Foundation Models, durchaus möglich sein könnte. Auch wenn dafür der innere Aufbau eines solchen Modells angepasst werden müsste, damit das Modell sich erfolgreich «fortbilden» kann. Wäre ein solches Modell tatsächlich in der Lage, eine Liste von wahrscheinlichen nächsten Worten zu erstellen, die nicht nur sprachlich korrekt, sondern auch inhaltlich passten, obwohl es nur sehr wenig Trainingsdaten gesehen hat, müsste es auch – Achtung, es wird wieder spekulativ – innere Gesetzmässigkeiten im Inhalt erkannt haben. Das könnte bedeuten, es hätte auch das menschliche Denken ein Stück weit entschlüsselt. Die Fähigkeit, die Wahrscheinlichkeiten von menschlicher Sprache zu erkennen, erschiene als Schlüssel für die Erledigung schwierigerer Aufgaben. Die Parallelen zur evolutionsbiologischen Entwicklung des Menschen wären offensichtlich. In der Schlussfolgerung hiesse das: Sprache und Verständnis/Intelligenz liessen sich nicht voneinander trennen. Wäre das tatsächlich der Fall, stünde dem Menschen die nächste «schwere Kränkung» bevor. In den nächsten ein bis zwei Jahren werden wir mehr wissen. ●

Christian R. Ulbrich
Burkhard Ringlein

Die Autoren

Christian R. Ulbrich ist Leiter der Forschungsstelle für Digitalisierung in Staat und Verwaltung an der Universität Basel. Er ist auch für das Wirtschaftsprüfungs- und Beratungsunternehmen PwC Schweiz tätig.

Burkhard Ringlein ist Postdoctoral Researcher bei IBM Research Europe und Vorstandsmitglied des liberalen Think Tanks LOAD e. V.