Generative KI: Nicht mehr als eine Blase?

ROMAN LEIPOLD
EXPERTE FÜR KÜNSTLICHE INTELLIGENZ BEI CHIP

Die großen Sprachmodelle von OpenAI und Co. produzieren nicht nur Texte und Bilder, sondern verbrennen auch viel Geld. Noch sprudeln die Investorengelder, doch die Kosten der KIs werden immer höher

Immer noch reiht OpenAI einen Superlativ an den anderen. In ihrer jüngsten Finanzierungsrunde hat die KI-Firma 6,6 Milliarden Dollar an Venture Capital eingesammelt. Noch nie haben Wagniskapitalgeber so viel Geld in ein einzelnes Unternehmen investiert. Mit einer Ausnahme, und auch da war OpenAI der große Profiteur: Anfang 2023 pumpte Microsoft zehn Milliarden Dollar in den Entwickler von ChatGPT, zu dem es aber schon seit 2019 als Hauptinvestor eine enge Geschäftsbeziehung unterhält.

Mit den Investitionen änderte sich auch die Bewertung von Open AI. Die Firma ist nun 157 Milliarden Dollar wert, doppelt so viel wie noch vor einem Jahr und mehr als Dax-Schwergewichte wie Siemens, Telekom oder Allianz.

Sprachmodelle: Teures Training

Das Kapital aus externen Quellen sprudelt, die Bewertung wächst in exponentiellem Tempo und die Nutzerzahlen kennen nur eine Richtung: nach oben. OpenAI badet im Erfolg – auf den ersten Blick. Bei genauerer Betrachtung wird nicht zuletzt durch die massive Fremdfinanzierung deutlich, dass der Investorenliebling unter einem massiven Kostenproblem leidet.

OpenAI braucht allein deshalb so viel Geld, weil der Betrieb von Large Language Models (LLMs) wie ChatGPT unglaublich teuer ist. Das beginnt beim Trainieren der Modelle mit riesigen Datenmengen. Für die parallele Verarbeitung der Daten braucht es spezielle Grafikchips (GPUs), auf deren leistungsstärkste Modelle Nvidia ein Quasimonopol besitzt.

Die führenden Entwickler von LLMs betreiben ihre KI mit Zehntausenden von GPUs. Laut einer Schätzung von Omdia Research hat Nvidia 2023 je 150.000 H100-GPUs an Microsoft und Meta ausgeliefert. Dieser Chip kostet zwischen 30.000 und 40.000 Euro pro Stück. Großartige Rabatte dürften selbst den Großkunden verwehrt bleiben, weil Nvidia die enorme Gesamtnachfrage immer noch nicht decken kann.

Das Nachfolgemodell, der GB200 wird auf 60.000 Euro taxiert; trotzdem hilft er den Großabnehmern, ihre Kosten zu senken, weil seine Effizienz den höheren Preis mehr als wett macht.

Nach Recherchen des Branchendienstes „The Information“ wird OpenAI allein in diesem Jahr für das Datentraining mit Hochleistungschips sieben Milliarden Dollar ausgeben. Hinzu kommen Ausgaben in Milliardenhöhe für den Alltagsbetrieb. Jede Anfrage eines Nutzers frisst viel mehr Rechenkapazität als zum Beispiel eine klassische Google-Suche. Nach aktuellen Berechnungen der Analysten von Bestbrokers führt dies zu einem zehn Mal höheren Stromverbrauch.

Als weiteres Problem gesellen sich die rasant wachsenden Personalkosten hinzu. Ähnlich wie bei den GPUs sorgt der Wettbewerb in Verbindung mit dem Fachkräftemangel für horrende Kosten. Laut „The Information“ gibt OpenAI 2024 für seine 1.700 Angestellten 1,7 Milliarden Dollar aus. Pro Mitarbeiter wendet die KI-Firma also im Durchschnitt eine Million Dollar pro Jahr auf. Weitere Steigerungen sind natürlich nicht ausgeschlossen.

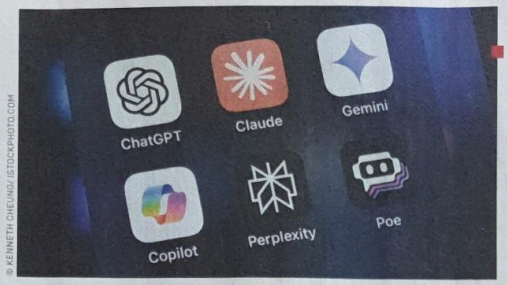
Angesichts dieser Belastungen kann ChatGPT nicht kostendeckend betrieben werden. Die „New York Times“ hatte Einsicht in interne Firmenunterlagen, die für 2024 Verluste in Höhe von rund fünf Milliarden US-Dollar prognostizieren. Der Umsatz soll sich im laufenden Jahr dagegen nur auf 3,7 Milliarden Dollar belaufen.

Im nächsten Jahr sollen den Unterlagen zufolge 11,6 Milliarden Dollar umgesetzt werden und in fünf Jahren will man mehr als 100 Milliarden Dollar erlösen. Erst dann rechnet OpenAI mit einem Gewinn, berichtet „The Information“. Erreicht werden sollen die Umsatzsprünge hauptsächlich durch Erhöhungen der Preise für Premiumversionen der LLMs.

Mehr Nutzer heißt: mehr Kosten

Dass diese Rechnung aufgeht, ist längst nicht ausgemacht. Selbst wenn die erhoffte Umsatzsteigerung eintritt, braucht OpenAI neues Kapital, da steigende Nutzerzahlen und wachsender Trainingsaufwand die Kosten weiter in die Höhe treiben.

Die Userzahlen haben im Frühjahr stark zugenommen, weil Open AI seit April die Nutzung ohne Registrierung erlaubt. So wurde im Juni mit 2,9 Milliarden Visits ein neuer Rekord verbucht. 200 Millionen Menschen greifen mindestens einmal im Monat auf ChatGPT zu. Damit hat sich die



© KENNETH CHEUNG/ISTOCKPHOTO.COM

Große Modelle mit kleinen Umsätzen
Die führenden Sprachmodelle wie ChatGPT haben mit wachsenden Kosten bei niedrigen Einnahmen zu kämpfen


© IMELJUDICE

„Echte Transformationen werden nicht schnell geschehen, nur wenige – wenn überhaupt – binnen zehn Jahren.“
Daron Acemoglu
WIRTSCHAFTS-NOBELPREISTRÄGER


© KIDFLY182

Goldman Sachs warnt vor KI-Hype
In einem Branchenreport weist die Investmentbank auf wachsende Finanzprobleme der Betreiber großer Sprachmodelle hin

Zahl der regelmäßigen Nutzer im Lauf des Jahres 2024 verdoppelt.

Die hohe Geldverbrennungsquote von OpenAI und seinen Wettbewerbern wirft die Frage auf, ob sich die Investitionen in generative KI (GenAI) je bezahlt machen. Die Investmentbank Goldman Sachs hat unter dem Titel „GenAI: Zu hohe Ausgaben, zu geringer Nutzen?“ einen Bericht veröffentlicht, in dem führende Ökonomen Zweifel anmelden.

Das 1-Billion-Dollar-Problem

Jim Covello, Chefanalyst von Goldman Sachs, glaubt, dass die Kosten für die Entwicklung und den Betrieb von GenAI nur zu einer angemessenen Rendite führen können, wenn die KI-Anwendungen äußerst komplexe und wichtige Probleme für die Unternehmen lösen. Covello: „Wir schätzen, dass allein der Aufbau der KI-Infrastruktur in den kommenden Jahren mehr als eine Billion Dollar kosten wird. Die entscheidende Frage lautet also: Welches 1-Billion-Dollar-Problem wird KI lösen?“

Laut Covello ist das Ersetzen von Niedriglohnjobs durch enorm teure Technologie im Grunde das genaue Gegenteil der früheren disruptiven Umwälzungen in der Digitalbranche: „Viele Menschen vergleichen heutige KI mit den Anfängen des Internets. Aber selbst in seinen Anfängen war das Internet eine kostengünstige Technologielösung, die es dem elektronischen Handel ermöglichte, teure etablierte Lösungen zu ersetzen. Amazon konnte Bücher zu niedrigeren Kosten als Barnes & Noble verkaufen, weil es keine kostspieligen Ladengeschäfte unterhalten musste.“

Diese Disruption halte bis heute an, etwa bei Taxi- und Chauffeurdiensten, die zunehmend durch Uber ersetzt würden.

Über die Frage, ob KI jemals die Versprechen einlösen werde, die viele Menschen heute in Euphorie versetzten, könne man geteilter Meinung sein, so Covello. Weniger umstritten sei aber der Punkt, dass KI-Technologien außergewöhnlich teuer seien: „Und um diese Kosten zu rechtfertigen, muss die Technologie in der Lage sein, komplexe Probleme zu lösen, für die sie nicht ausgelegt ist.“

Auch Daron Acemoglu, Wirtschaftsprofessor am Massachusetts Institute of Technology in Boston, wurde von Goldman Sachs zu den Zukunftsaussichten von GenAI befragt. Der frischgebackene Nobelpreisträger erforscht die historischen Ursprünge von Wohlstand und Armut. In seinem aktuellen Buch „Macht und Fortschritt“ untersucht er die Auswirkungen neuer Technologien auf Wachstum, Beschäftigung und Ungleichheit.

Das Potenzial Künstlicher Intelligenz, neue Produkte, Materialien und Prozesse zu entwickeln, zieht der Starökonom nicht

OpenAI könnte 2024 einen Verlust von 5 Milliarden Dollar anhäufen

in Zweifel. „Doch angesichts der Ausrichtung und der Architektur von generativer KI werden diese wirklich transformativen Veränderungen nicht schnell passieren und nur wenige – wenn überhaupt – innerhalb der nächsten zehn Jahre.“

Acemoglu ist in einer Studie zu dem Schluss gekommen, dass GenAI allenfalls fünf Prozent der beruflichen Tätigkeiten

ChatGPT lockt das große Geld an

Seit dem Durchbruch der generativen KI vor zwei Jahren pumpen Investoren Rekordsummen in OpenAI und Co

QUELLE: 2024 AI INDEX REPORT

QUELLE: THE INFORMATION

substanziell beeinflusst: „Viele Aufgaben, die Menschen derzeit ausführen, zum Beispiel in den Bereichen Transport, Fertigung oder Bergbau sind vielschichtig und erfordern eine Interaktion mit der realen Welt – eine Fähigkeit, mit der KI in absehbarer Zeit nicht wirklich mithalten kann.“.

Daher, so Acemoglu, würden sich die größten Auswirkungen der Technologie in den nächsten Jahren wahrscheinlich um reine Denkaufgaben drehen, die in Anzahl und Umfang zwar nicht trivial, aber eben auch nicht wirklich mächtig seien.

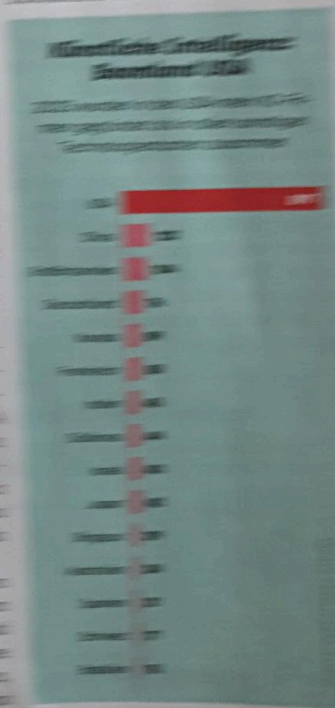
Acemoglu weiß natürlich, dass KI mit der Zeit besser wird, er glaubt aber, dass große Datenmengen und Rechenleistung allein die Technologie nicht entscheidend voranbringen: „Viele Leute in der Branche scheinen an eine Art Skalierungsgesetz zu glauben, wonach eine Verdoppelung der Datenmenge und der Rechenkapazität die Fähigkeit von KI-Modellen verdoppelt. Ich würde diese Ansicht jedoch in mehrfacher Hinsicht in Frage stellen. Was bedeutet es,

© WIKIMEDIA COMMONS

Goldene Ära für amerikanische Technologiefirmen
Mit der Machtübernahme durch Donald Trump und J.D. Vance dürften die Zeiten staatlicher Regulierung erst mal vorbei sein

© OFFICE OF GOVERNOR WALZ & LT. GOVERNOR FLANAGAN

Keine Mehrheit für die Regulierung von KI
Kamala Harris und Tim Walz scheiterten mit ihren vorsichtigen Bemühungen, die Entwicklung großer KI-Modelle zu kontrollieren



strengerer Kontrolle unterliegen, ist eine überfällige Maßnahme, die sich andere Länder zum Vorbild nehmen sollten.

In den USA wird das erst einmal nicht passieren. Nach dem Wahlsieg von Donald Trump schossen die Börsenkurse der KI-Firmen in die Höhe. Elon Musk, ohnehin schon reichster Mann der Welt, durfte sich über 21 Milliarden Dollar freuen, die er über Nacht, aber nicht im Schlaf verdiente. Er verbrachte den Wahlabend nämlich in Florida bei seinem Buddy Donald Trump.

Der Präsident will den gebürtigen Südafrikaner als „Effizienz-Zar" in sein Regierungsteam zu holen. So könnte Musk zum Beispiel Genehmigungsverfahren für die von ihm vorangetriebenen digitalen Gehirnimplantate „entbürokratisieren" und Regierungsaufträge für sein Raumfahrtunternehmen SpaceX anleiern. Auch sein Start-up xAI müsste sich wohl nicht länger mit der KI-Gesetzgebung herumschlagen, die das Duo Biden und Harris in seiner Legislaturperiode auf den Weg gebracht hat. Ähnliches gilt für die Auflagen, an die sich Tesla bisher bei seinen Experimenten mit autonomer Fahrtechnik halten musste. 21 Milliarden Dollar Gewinn und ein Freifahrtschein für seine Firmenprojekte – die rund 140 Millionen Dollar, mit denen Musk Trump im Wahlkampf unterstützte, haben sich gelohnt.

Aber auch die Betreiber der KI-Datenzentren wie OpenAI, Microsoft, Amazon und Google profitieren von der „Goldenen Ära", die Trump ausgerufen hat, weil es weniger Auflagen, Regeln und Gesetze geben wird. Das betrifft nicht nur die Datenzentren und die sündhaft teuren LLMs, sondern auch die Stromversorgung. Die Techkonzerne suchen in den USA den direkten Anschluss an Atomkraftwerke, um die Versorgung ihrer energiefressenden

„Die Frage lautet: Welches 1-Billion-Dollar-Problem wird KI lösen?"
Jim Covello
AKTIENANALYST GOLDMAN SACHS

Kreationen mit Strom dauerhaft abzusichern. Der bis vor Kurzem noch massive Widerstand von Behörden wie der Federal Energy Regulatory Commission dürfte unter dem Duo Trump/Musk schnell erlahmen. Der KI-Experte Gregory Allen, der unter Trump und Biden für das Verteidigungsministerium gearbeitet hat, prophezeit, dass der Präsident den Bau von AKWs komplett deregulieren wird.

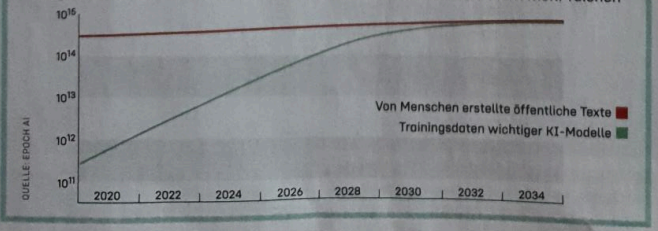
Selbst wenn die finanziellen Aussichten der KI-Anbieter unter Trump etwas rosiger aussehen, bleibt die 1-Billion-Dollar-Frage von Jim Covello offen. Daron Acemoglu erinnert daran, dass ChatGPT im Grunde nicht viel mehr als ein „stochastischer Papagei" sei, der aufgrund von Wahrscheinlichkeiten das jeweils nächste Wort in einem Text bestimmt: „Große Sprachmodelle haben sich als beeindruckender erwiesen, als viele Menschen erwartet haben. Aber es ist immer noch ein großer Vertrauensvorschuss erforderlich, um zu glauben, dass die Architektur der Vorhersage des nächsten Wortes in einem Satz Fähigkeiten erreichen wird, die so intelligent sind wie HAL 9000 in '2001: A Space Odyssee'. Es ist so gut wie sicher, dass die derzeitigen Modelle innerhalb der nächsten zehn Jahre eine solche Leistung nicht annähernd vollbringen werden."

700.000 Dollar
zahlt OpenAI pro Tag, um die Server für ChatGPT zu betreiben.

7 Milliarden Dollar
kostet OpenAI 2024 das Datentraining seiner KI-Modelle.

Der Künstlichen Intelligenz gehen die Daten aus

Um sich zu verbessern, brauchen LLMs hochwertige Trainingsdaten. Doch der dafür nötige Nachschub an Texten aus Menschenhand wird schon bald nicht mehr reichen



Dass es mit der Präzision von LLMs noch nicht weit her ist, hat OpenAI mit einem selbst entwickelten Benchmark gerade unter Beweis gestellt. Für den Test „SimpleQA" haben KI-Trainer über 4.000 Fragen erstellt, wobei es immer nur eine korrekte Antwort gibt, die sich auch im Lauf der Zeit nicht verändert.

Für LLMs ist Wissen Glückssache

Wie ein klassisches TV-Quiz deckt „SimpleQA" alle Themenbereiche des Lebens ab, fordert aber zum Teil extremes Detailwissen. Eine Frage geht zum Beispiel bis ins Jahr 1946 zurück: „Wer gewann das erste Finale des Hessen-Pokals im Fußball?". Die richtige Antwort „Eintracht Frankfurt" gehört sicher nicht zur Allgemeinbildung von US-Bürgern, aber der Test sollte die LLMs ja auch herausfordern. Was auch gelang: Alle Modelle, nicht nur ChatGPT, schnitten schlecht ab.

OpenAIs o1-Preview, die erst im Oktober 2024 veröffentlicht wurde, gewann den Test mit der dürftigen Trefferquote von 42,7 Prozent. Claude-3,5-Sonnet, das aktuelle Konkurrenzmodell von Anthropic lag nur bei 28,9 Prozent aller Fragen richtig. Das bedeutet, dass die Antworten, welche die beiden führenden Modelle im Test geben, öfter falsch als richtig sind.

Für Acemoglu bestätigen diese Schwächen, dass sich KIs, die so allgegenwärtig und mächtig sind, als gefährlich herausstellen könnten. Deshalb sei es wichtig, langsamer vorzugehen. Man müsse „dem Hype widerstehen und einen vorsichtigeren Ansatz wählen, zu dem auch bessere Regulierungsinstrumente gehören".

Generell fordert der Nobelpreisträger mehr Bedacht: „Das Risiko, dass unsere Kinder oder Enkelkinder uns im Jahr 2074 vorwerfen, wir hätten uns 2024 auf Kosten des Wachstums zu langsam bewegt, scheint weitaus geringer zu sein als das Risiko, dass wir uns zu schnell bewegen und dabei Institutionen, die Demokratie und noch mehr zerstören."