Roland Kunz

Transformer – Textgenerierung, Industrieanwendungen und Grenzen

Abstract: Transformer based neuronal networks, engine of the Generative AI wave, promise a new and more powerful method of generating text, images, and sound. All over the industry, companies are evaluating the proper use of this new technology and the associated risks and benefits. This paper will discuss the current industry trends and observations, as well as the limitations of that architectural approach. Considerations on where to engage on this journey are discussed as well as some practical limitations.

1 Einführung

Anwendungen im Bereich Künstlicher Intelligenz (KI),[1] welche auf dem Konzept der sogenannten General Pretrained Transformers (GPT) beruhen,[2] haben in der Industrie eine breite Resonanz verursacht,[3] ermöglichen sie doch eine wesentlich breitere Anwendung, als vorher existierende Lösungen. Diese als generative KI Anwendungen (Generative AI) bezeichneten Systeme sind insbesondere durch Lösungen wie Chat-GPT bekannt geworden.[4] Als Beispiel für die Relevanz für Industrie und bei der Suche nach den richtigen Ansätzen sei hier eine Gartner-Studie zu erwähnen, in der weltweit Kunden zu diesem Thema befragt wurden. Hier zeigt sich, dass bereits fast die Hälfte ihre KI-Investitionen allein aufgrund des Potenzials von Chat-GPT und generativen KI-Anwendungen erhöht haben. In drei Jahren – so diese Gartner Untersuchung – wird erwartet, dass mehr als 60 % der Anwendungen wie Telefon-, Web- oder Windows-Apps automatisch durch KI-Codierung, Programmierung und generative Funktionen erstellt werden.[5] Das Versprechen dieser

1 Richard Lackes, „Künstliche Intelligenz (KI)“, *Gabler Wirtschaftslexikon*, in [www.wirtschaftslexikon.gabler.de/definition/kuenstliche-intelligenz-ki-40285] (Zugriff: 04.10.2023).

2 „GPT“, *cambridge dictionary*, in [www.dictionary.cambridge.org/dictionary/english/gpt] (Zugriff: 04.10.2023).

3 „Cathie Wood on Deflation Risk, Tech Stocks and Bitcoin“, *Bloomberg*, in [https://t1p.de/34lwd] (Zugriff: 01.10.2023).

4 OpenAI, *ChatGPT*, in [www.openai.com/chatgpt] (Zugriff: 04.10.2023).

5 Medha, „The Generative AI Landscape: Where We Stand and Where We're Headed“, *Fireflies.ai Blog*, 2023, in [www.t1p.de/shpvm] (Zugriff: 04.10.2023).

 https://doi.org/10.1515/9783111351490-028

Lösungen für die Industrie liegt hierbei in der Automatisierung und Produktivitätssteigerung, nach der Unternehmen suchen. Ebendies wurde auch in einer weiteren Studie von Bloomberg und ARK Investment bestätigt.[6] Diese Studie sagt, dass Unternehmen mit vielen qualitativ hochwertigen Daten besonderen Nutzen und Wettbewerbsvorteile aus dieser Technologie ziehen werden können. Weiterhin ist es für Unternehmen essentiell, die Balance zwischen dem Zeitpunkt des Einstiegs und der Größe der Investition auf der einen und der Lernkurve durch die adaptive Einführung in der Organisation auf der anderen Seite abzuwägen. Insbesondere seit und durch die Adaption von generativer KI und trotz diverser externer Einflüsse wie Inflation und politische Konflikte ist dieser Markt ein sehr schnell wachsender.

Die Abschätzung, wie stark dieser Markt wachsen wird und was Führungskräfte in der Industrie sagen, ist hier aufgelistet:

- KI wird 15.7 Milliarden Dollar zur Weltwirtschaft im Jahr 2030 beitragen
- 45 % aller Führungskräfte erhöhen ihre Investitionen, auch aufgrund von ChatGPT[7]
- 73 % aller CTOs sehen die Rolle von KI innerhalb ihrer Unternehmen in den nächsten 2 Jahren als essentiell an.
- 75 % aller großen Unternehmen werden KI nutzen, um ihre Effizienz und Qualität zu steigern.[8]

Auch eigene Untersuchungen von Dell Technologies bei 200 Führungskräften innerhalb der IT zeigen, dass 91 % generative KI zu Hause und 71 % diese auf der Arbeit nutzen.[9]

2 Anwendungsbeispiele in der Industrie

Basierend auf der deutlich gestiegenen Leistungsfähigkeit heutiger Computer und der Beschleunigung von KI Anwendungen,[10] sind auch die Anwendungsbeispiele innerhalb vieler Unternehmen und Branchen deutlich gestiegen.

---

6 Chaim Haas/Alyssa Gilmore, „Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance", *Bloomberg,* 2023, in [https://t1p.de/oy833] (Zugriff: 02.10.2023).

7 Esther Shein, *Gartner: ChatGPT interest boosts generative AI investments,* 2023, in [www.techrepublic.com/article/gartner-executives-chatgpt-investments/] (Zugriff: 02.10.2023).

8 IDC FutureScape, *Worldwide AI and Automation 2023 Predictions,* 2022.

9 Dell Untersuchungen, Februar 2023, intern.

10 JP Mangalindan, *A timeline of computing power,* in [www.money.cnn.com/interactive/technology/computing-power-timeline/] (Zugriff: 04.10.2023).

Stand dieser Publikation sind bereits einige Unternehmen erfolgreich damit, generative KI in ihren eigenen Softwareprodukten einzusetzen,[11] aber auch wenn generative KI genutzt werden kann, um menschliche Interaktion und schlussendlich Arbeitskraft einzusparen, geht die Tendenz der meisten Analysten in die Richtung, sie dort einzusetzen, wo diese KI repetitive Aufgaben identifizieren und ersetzen kann. Diese umfassen die im folgenden aufgezeigten Felder:

- Digitale Assistenten im Bereich Vertrieb, Sicherheit, Infrastruktur, Retail und Business Oprations. Diese helfen Prozesse zu beschleunigen. Dies wird insbesondere durch den schnelleren Zugriff auf die richtigen Daten und die Möglichkeit, neue Daten zu trainieren erreicht. Beispiele seien hier Suchanfragen, Empfehlungen und Kundenvorlieben.
- Entwickler sind in der Lage, effizienter Code oder User Interface (UI) und User Experience (UX) Designs zu erstellen. Dies erhöht nicht nur die Produktivität, sondern hilft, früher in das Produktdesign und die Realisierung einsteigen zu können. Zusätzlich sinkt die Anzahl der Fehler in automatisch generiertem Code, was ein wesentlicher Aspekt für mehr Sicherheit ist.
- Vertrieb und Marketing können über die Erstellung von kreativen Inhalten und Vertriebsskripten eine hohe Bandbreite an Kommunikationsmitteln nutzen, um mit ihren Kunden in Kontakt zu kommen. Dies beinhaltet nicht nur Text, sondern auch Audio, Video, Bilder, Podcasts, Videospiele und andere Anwendungen. Hierdurch wird die Skalierung als auch die Qualität erhöht.
- Für Forschung, Wissenschaft und Lehre kann es einen Vorteil bei der Erstellung von technischen Dokumenten im Hinblick auf Genauigkeit und Geschwindigkeit der Veröffentlichung bis hin zum Gutachten mit sich bringen.

Auch für diese Publikation wurde der letzte Punkt der vorherigen Aufzählung anhand eines Experiments mit ChatGPT durchgeführt. Hierbei wurde dem Tool die Frage nach Anwendungsbeispielen im Bereich Finanzdienstleistungen genannt, und als Antwort die in folgender Abbildung 1 dargestellten Ergebnisse geliefert.

11 Adobe, „Your imagination's new best friend", *Generative AI,* in [www.t1p.de/gell6] (Zugriff: 04.10.2023).

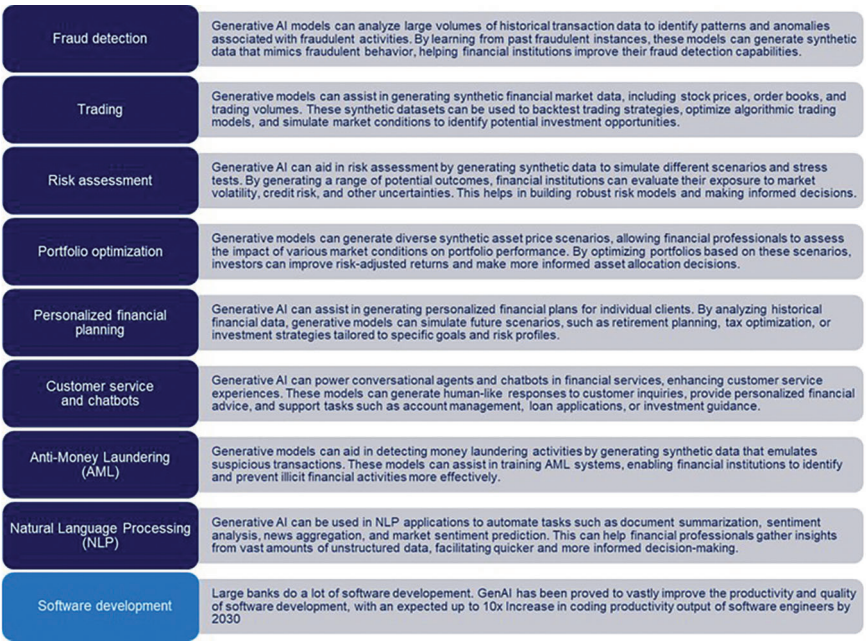

Abbildung 1: Ergebnisse von ChatGPT auf die Frage nach Use Cases für die Finanzbranche, durchgeführt von Dell Technologies im Mai 2023 auf englischer Sprache.[12]

3 Transformer – Basis und Status

Die Möglichkeiten solcher Lösungen auf Basis von generativer KI und Transformern sind also groß – wie aber sieht es mit der Handhabbarkeit und der Nutzung aus? Transformer, die Basis von solchen Modellen, sind keine einfachen Konstrukte, sondern komplexe Abfolgen von Algorithmen, die dem Ziel dienen, eine statistisch sinnvolle Abfolge von Worttokens zu generieren. Hierbei werden Bausteine wie Satz- und Wordklassifizierung, die Beantwortung von Fragen oder das Ausfüllen von fehlenden Satzteilen benutzt ebenso wie die komplette Synthese eines Textes, dessen Übersetzung und Zusammenfassung. Der Transformer ist eine Art Motor der eigentlichen Anwendung, und daher streben viele kommerzielle und nicht-kommerzielle Unternehmen und Institute nach dem besten Basismodell. Eine von der

12 Prompt: „List the most important use cases for the financial services industry for generative AI“, *Open AI*, in [https://openai.com/chatgpt] (Zugriff: 13.05.2023).

Website LifeArchitect.ai[13] veröffentlichte Zusammenfassung der aktuellen Modelle, zeigt hierbei, dass ein Streben nach dem größten Modell im Gange ist, wie in Abbildung 2 zu sehen ist. Größeren Modellen wird hierbei eine bessere Fähigkeit zur Abstraktion zugeschrieben.

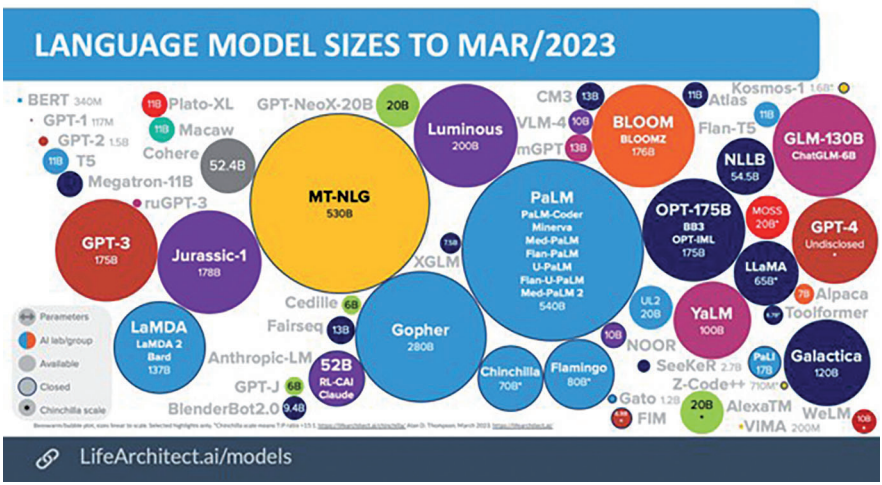

Abbildung 2: Größe aktueller Transformer Modelle. Die Zahl innerhalb der Kreise (zum Beispiel 530B) beschreibt die Anzahl der zu trainierenden Parameter (Billons = Milliarden).

4 Industrieüberlegungen

Wie im vorherigen Abschnitt gezeigt, setzt die Nutzung von generativer KI unter Verwendung von Transformern ein gewisses Maß an Planung voraus. Insbesondere der Grad der Fertigungstiefe, also wie viel die Kunden an eigener Leistung einbringen wollen, ist hierbei von Relevanz. Prinzipiell kann man hierbei drei verschiedene Ansätze unterscheiden:

1. Kunden konsumieren ein Angebot in einer Cloud oder betreiben das Angebot in Form eines fertigen Produktes bei sich in der eigenen Umgebung. Dieser „Inferenz“ genannte Teil des Lebenszyklus von KI erfordert die geringsten Anforderungen an die eigene Infrastruktur. Lediglich die Art, den Dienst zu konsumieren, ist festzulegen. Sei es als „Software as a Service“, wie zum Bei-

13 Alan D. Thompson, *Interne Sprachmodelle (von GPT-4 bis PaLM),* in [www.lifearchitect.ai/models] (Zugriff: 06.10.2023).

spiel durch Nutzung der APIs oder der Oberfläche von ChatGPT, als Instanz einer großen Cloud, zum Beispiel Microsoft Azure oder im eigenen Betrieb. Hierbei müssen die bereits früher trainierten Daten verwendet werden. Nicht selten nutzen diese Modelle den weiteren Input der Nutzer für weitere Trainings. In Bezug auf eigene Daten muss hier also sehr sorgsam abgewogen werden, da unter Umständen Unternehmenswissen Teil der neuen Trainingsdaten werden können, und somit prinzipiell anderen Unternehmen zur Verfügung stehen könnten.

2. Kunden trainieren ihr eigenes Transformer basierendes Large Language Modell mit ihren eigenen Daten und potentiell öffentlich verfügbaren allgemeinen Daten. Hierzu besteht neben der Notwendigkeit, hinreichend viele eigene Daten in guter Qualität zu haben, auch die Anforderung an eine sehr leistungsfähige Infrastruktur. Diese Investition ist gründlich zu bedenken, bevor ein eigenes System trainiert wird. Ein solches System stellt allerdings prinzipiell die beste Möglichkeit dar, externen Bias zu vermeiden und eigene, private Daten zu schützen. Ein so trainiertes Modell kann dann natürlich zur Inferenz genutzt werden.
3. Kunden nehmen ein existierendes Modell, welches mit generischen Daten trainiert wurde, und trainieren es mit eigenen Daten weiter. Aus dem bis dahin öffentlichen Modell wird ein privates – mit dem Schutz der eigenen Daten, aber der Unsicherheit in Bezug auf möglichen Bias der Ursprungsdaten.

4.1 Herausforderungen der Technologie

GPT-3 ist ein gut dokumentiertes Modell[14] und soll hier als Beispiel dienen, was für jeden der oben genannten Möglichkeiten (Training, Finetuning, Inferenz) notwendig wäre, wenn ein Unternehmen ebendieses nutzen will. Unter der Annahme hinreichend guter Daten und dem Versuch, solch ein Modell selber zu trainieren, bedeutete dies:

Pro Parameter des Netzwerkes und zu lernendem Token müssen 6–8 flops (floating point operations per second) an Rechenleistung angesetzt werden. Die Anzahl der Rechenschritte ist proportional zu der Anzahl der Parameter und der zu trainierenden Token. Bei einer Größe von 175 Milliarden Parametern und 300 Milliarden Tokens[15] ergibt sich also ein theoretischer Rechenbedarf von

$175.10^{e9} * 300.10^{e9} * 8 \text{ flops} = 0.4.10^{e24} \text{ flops}$

14 Min Zhang/Juntao Li, „A commentary of GPT-3“, *MIT Technology Review*, 2021, in [https://t1p.de/s4q6o] (Zugriff: 06.10.2023).

15 Wikipedia, „GPT-3“, in [https://en.wikipedia.org/wiki/GPT-3] (Zugriff 12.10.23).

Zum Berechnen dieser Parameter werden aktuell vor allem NVIDIA GPUs des Typs A100 oder H100 verwendet. Die A100 hat dabei im Bereich Bfloat16 eine theoretische Leistung von 312 tflops/s.[16] Die tatsächliche Leistung kann hier nur geschätzt werden, da Benchmarks nicht mit dem wirklichen Training gleichzusetzen sind. Bei 40–50 % Effektivität wird also ein Wert von etwa 125 tflop/s angesetzt. Eine einzelne Karte würde also

$0.4.10^{e24} / 125.10^{e12} = 3.2.10^{e9}$ s

Rechenzeit benötigen, was umgerechnet 37037 Tage sind. Unter der Annahme linearer Skalierung würde ein System bestehend aus 256 Knoten mit je 8 solcher GPUs ein solches Modell in 20 Tagen trainieren. Nutzt man die leistungsfähigere H100 GPU, würden statt 256 Knoten nur noch 80 benötigt werden, aber selbst diese Anzahl stellt in Bezug auf die Infrastruktur einen erheblichen Aufwand dar.

Ein typisches System für solch ein Training ist der Dell PowerEdge XE9680,[17] welcher 8 GPUs vom Typ H100 zur Verfügung stellt, 6 Höheneinheiten in einem Standard Rack belegt und je nach restlicher Ausstattung bis zu 11.5 kW Stromaufnahme hat. Ein durchschnittliches Rack in einem typischen Europa liefert 15 kW,[18] so dass hier bereits für viele Kunden eine Herausforderung besteht.

Betrachten wir hierzu beispielhaft die Energiekosten, die der Betrieb eines solchen Systems benötigt. In der EU lag der durchschnittliche Strompreis im Jahr 2023 bei ungefähr 30ct/kWh[19] und unter Verwendung eines Effizienzfaktors von 1,5 (PUE für luftgekühlte Rechenzentren)[20] kommt man zu Stromkosten von

11.5 kW/h * 80 * 1,5 * 0.30 = 28k€

Je nach Land und Jahreszeit sind das durchschnittlich 300geqCO^2/kWh CO^2 Emission bei einer Bandbreite von 50–800geqCO^2/kWh.

Diese Herausforderungen lassen erwarten, dass die meisten Kunden eher Inferenz oder fortgesetztes Training (Finetuning) nutzen werden.

Jedoch ist nicht nur die schiere Größe eines LLMs ein Erfolgsgarant. Mittlerweile gibt es viele Bestrebungen, kleinere, optimierte Systeme zu konstruieren. Ein Beispiel ist das spezialisierte System BloombergGTP.[21]

---

16 NVIDIA, in [https://t1p.de/so7qo] (Zugriff 12.10.2023).

17 DELL, in [https://www.dell.com/en-us/shop/ipovw/poweredge-xe9680] (Zugriff 12.10.2023).

18 Bernhard Rohleder, *Mehr Daten – mehr Strom? Wie sich Rechenzentren in Deutschland entwickeln,* 2022, in [https://t1p.de/dt3qb] (Zugriff: 06.10.2023).

19 Stefanie Schäffer, *Strompreise Europa 2023 – Was kostet Strom in der EU?,* in [www.energiemarie.de/strompreis/europa] (Zugriff: 06.10.2023).

20 *PUE Wert – der Indikator für die Effizienz eines Rechenzentrums,* 2023, in [https://www.t1p.de/3eu2m] (Zugriff: 06.10.2023).

21 Katharina Buchholz, *ONE MILLION USERS. Threads Shoots Past One Million User Mark at Lightning Speed,* 2023, in [www.t1p.de/z7xnw] (Zugriff: 09.09.2023).

Dieses Modell ist mit 50 Milliarden Parametern deutlich kleiner als die stetig wachsenden Modelle etwa von OpenAI, es wurde auch mit spezialisierten Daten, bestehend aus 700 Milliarden Tokens englischer Finanzpublikationen und allgemeiner Information trainiert. Im Vergleich zu GPT-3 ein deutlich kleineres Modell, aber mit mehr Tokens.

Insgesamt ist eine Diversifizierung von Modellen durch die OpenSource Community zu erkennen, so dass kleinere, spezialisiertere Modell zu erwarten sind.

4.2 Spezialisierung und Voraussetzungen

Neben dem Betrieb eines solchen Systems ist eine weitere Herausforderung die Verfügbarkeit von Spezialisten, um ein solches System den eigenen Bedürfnissen anzupassen. Ein Vortrainiertes LLM, was zum Beispiel mit 1 Milliarde Tokens trainiert wurde, muss für die eigene Anpassung mit einigen eigenen Daten trainiert werden. Dies geschieht oft durch Prompt und Response Trainings, die überwacht werden anhand von Demonstrationsdaten. Ein so trainiertes System muss dann Vergleichsdatensätze durchlaufen und wird mit Reinforcement Learning weiter abgestimmt.

Weiterhin muss gewährleistet werden, dass das System keine sachlich falschen, oder unethischen Antworten gibt. Dies wird in der Regel über so genannte Guardrails erreicht, die dem System die Prompts vorgeben und die Nutzereingaben prüfen. Auch ist der Bias eines solchen Modells immer ein Punkt, der mit betrachtet werden muss.

5 Fazit

Der Einsatz von Large Language Modellen auf Basis der Transformer Architektur ist für jedes Unternehmen derzeit ein Thema in Bezug auf Nutzen, Wettbewerbsvorteile und Investition. Diese disruptive Technologie bietet Chancen im Wettbewerb, aber auch Risiken, die es abzuwägen gilt.