

Large Language Models: Eine Bestandsaufnahme

Whitepaper (Oktober 2024)

© Deepsearch GmbH

Inhaltsverzeichnis

| | |
|---|-----------|
| Zusammenfassung | 3 |
| Einleitung | 4 |
| Technologieüberblick | 5 |
| Generative Pretrained Transformer (GPT) | 5 |
| Recurrent Neural Networks (RNN) | 5 |
| Convolutional Neural Networks (CNN) | 6 |
| Reinforcement Learning | 6 |
| Neurosymbolic AI | 6 |
| Large Language Models (LLMs) | 7 |
| Zuverlässigkeit und Relevanz | 8 |
| Abstraktionsvermögen und „Reasoning“ | 9 |
| Transparenz | 12 |
| Planungsfähigkeit | 13 |
| Vertrautheit vs. Komplexität | 14 |
| Conclusio | 16 |
| Quellen | 17 |
| Abbildungsverzeichnis | 18 |



Zusammenfassung

Diese systematische Bestandsaufnahme analysiert den aktuellen Entwicklungsstand von Large Language Models (LLMs) und deren Grenzen im unternehmerischen Kontext. Die Untersuchung fokussiert vier zentrale Aspekte: Zuverlässigkeit und Relevanz, Abstraktionsvermögen und Reasoning, Transparenz sowie Planungsfähigkeit. Die empirische Analyse belegt, dass LLMs zwar beeindruckendes Pattern-Matching auf hohem Niveau demonstrieren, jedoch bis dato keine echte Intelligenz entwickeln.

Während LLMs in ihren Kernbereichen wie Textgenerierung, Zusammenfassung und Übersetzung überzeugen, zeigen sich bei der Automatisierung komplexer Geschäftsprozesse signifikante Limitationen.

Es wird hervorgehoben, dass LLMs, die auf der Generative Pretrained Transformers (GPT) Architektur basieren, nur eine von mehreren KI-Technologien darstellen. Je nach Anwendungsfall eignen sich alternative Ansätze oder eine Kombination diverser Ansätze besser. Diese werden in einer Übersicht umrissen.

Basierend auf aktueller Primärliteratur zu LLMs werden vier konkrete Handlungsempfehlungen entwickelt:

1. Der gezielte, anwendungsspezifische Einsatz von KI-Technologien
2. Das Verständnis von LLMs als Baustein statt Komplettlösung
3. Die Entwicklung flexibler, robuster Systemarchitekturen
4. Die Implementierung einer Microservice-Architektur mit austauschbaren Komponenten

Diese Maßnahmen ermöglichen eine effektive Integration von KI-Technologien in bestehende Unternehmensarchitekturen. Die Microservice-Architektur gewährleistet dabei die notwendige Flexibilität, um mit der hohen Dynamik der KI-Forschung und -Entwicklung Schritt zu halten und neue Technologien nahtlos integrieren zu können.



Einleitung

Diese systematische Bestandsaufnahme analysiert den aktuellen Forschungsstand zum unternehmerischen Einsatz Künstlicher Intelligenz (KI), mit besonderem Fokus auf Large Language Models. Die Untersuchung basiert auf sorgfältig ausgewählter Primärliteratur aus wissenschaftlich fundierten Quellen. Sämtliche verwendeten Quellen sind sowohl im Fließtext als auch im Literaturverzeichnis vollständig dokumentiert.

Künstliche Intelligenz wird gemäß der Definition von Gartner verstanden als: "Artificial Intelligence (AI) applies advanced analysis and logic-based techniques — including machine learning (ML) — to interpret events, support and automate decisions, deliver experiences and take actions." (Mullen, 2024).

Generative KI (GenAI) ist ein spezifischer Teilbereich der Künstlichen Intelligenz, der seit der Veröffentlichung von ChatGPT im November 2022 die Aufmerksamkeit von Entscheidungsträgern weltweit auf sich zieht.

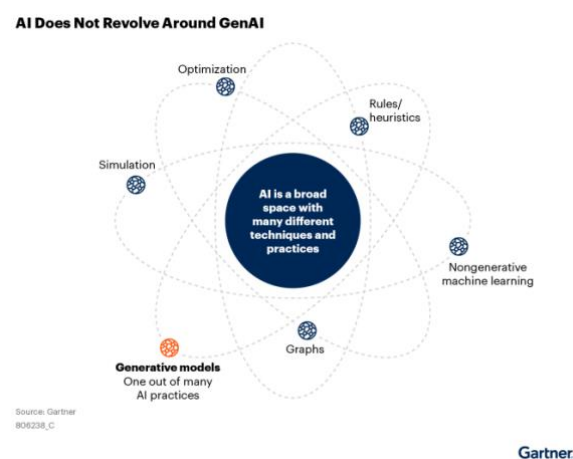


Abbildung 1: GenAI ist ein Teilbereich von KI

Large Language Models (LLMs) zeichnen sich durch ihre Fähigkeit aus, Text auf nahezu menschlichem Niveau zu generieren. Eine differenzierte Betrachtung ihrer Einsatzmöglichkeiten und Limitationen ist für eine effektive Implementierung dieser Technologie unverzichtbar.

Die sachliche Analyse der zugrundeliegenden Technologie wird durch den Marketing-Hype der weltweit führenden IT-Unternehmen erschwert. Wie Ramos et al. (2024) feststellen: "IT leaders struggle in understanding when and when not to apply generative AI (GenAI) for their use cases. The hype surrounding GenAI can lead them to apply it where it is not a good fit, increasing the risk of higher complexity and failure in their AI projects."



Dieses Dokument zielt darauf ab, technischen Entscheidungsträgern eine fundierte Orientierung zu bieten und sie bei der Entwicklung sachlich begründeter Einschätzungen zu unterstützen.

Technologieüberblick

In Folge wird eine Auswahl relevanter Techniken kurz erklärt, um einen Überblick zu ermöglichen und das Bewusstsein zu schärfen, dass es nicht nur LLMs gibt, sondern je nach Anwendungsfall andere Techniken besser geeignet sind. Jede Technologie hat ihre spezifischen Stärken und Schwächen, weshalb die Wahl der richtigen Technologie für den jeweiligen Anwendungsfall entscheidend für den Erfolg eines KI-Projekts ist.

Generative Pretrained Transformer (GPT)

Optimal für Sprachverarbeitung und -generierung, da sie durch ihre Transformer-Architektur komplexe sprachliche Zusammenhänge über lange Sequenzen hinweg erfassen können. Ein Schlüsselement ist dabei der Selbstaufmerksamkeitsmechanismus (Self-Attention), der es dem Modell ermöglicht, die Bedeutung jedes Wortes im Kontext aller anderen Wörter zu verstehen.

Beispiel: Ein GPT-System kann aus einer kurzen Aufgabenstellung wie "Schreibe eine Email an einen Kunden zur Terminverschiebung" eine vollständige, kontextgerechte und professionelle Email generieren.

Ungeeignet für: Echtzeitentscheidungen im autonomen Fahren, da GPT-Modelle zu langsam in der Inferenz sind und nicht für die direkte Verarbeitung von Sensordaten konzipiert wurden.

Recurrent Neural Networks (RNN)

Besonders effektiv für sequenzielle Daten wie Zeitreihen oder Spracherkennung, weil sie durch ihre rekurrente Struktur zeitliche Abhängigkeiten in den Daten berücksichtigen. Moderne Varianten wie LSTM (Long Short-Term Memory) lösen dabei das Problem des verschwindenden Gradienten durch spezielle Gedächtniszellen.

Beispiel: Ein RNN kann den Stromverbrauch eines Haushalts über die letzten Wochen analysieren und daraus präzise Vorhersagen für den Verbrauch der kommenden Tage treffen.

Ungeeignet für: Komplexe Bildverarbeitung, da RNNs die räumlichen Beziehungen in Bildern nicht effizient verarbeiten können.



Convolutional Neural Networks (CNN)

Hervorragend geeignet für Bildverarbeitung und Mustererkennung, da ihre Faltungsschichten räumliche Hierarchien in visuellen Daten effizient erfassen können. Die Faltungsoperation ermöglicht dabei die automatische Extraktion von Features verschiedener Abstraktionsebenen, von einfachen Kanten bis zu komplexen Objektstrukturen.

Beispiel: Ein CNN kann in medizinischen Röntgenbildern automatisch potenzielle Anzeichen für Lungenkrebs erkennen und Radiologen bei der Diagnose unterstützen.

Ungeeignet für: Verarbeitung von Texten variabler Länge, da CNNs nicht gut mit sequenziellen Daten unterschiedlicher Länge umgehen können.

Reinforcement Learning

Ideal für Entscheidungsfindung in dynamischen Umgebungen wie Robotersteuerung oder Spielestrategien, weil das System durch Trial-and-Error optimal handeln lernt. Die Verwendung von Belohnungssignalen (Rewards) und Strategien wie Q-Learning ermöglicht es dem System, langfristig optimale Aktionssequenzen zu entwickeln.

Beispiel: Ein Roboterarm in einer Fabrik lernt durch Reinforcement Learning, wie er Objekte unterschiedlicher Form und Größe präzise greifen und platzieren kann, indem er aus erfolgreichen und fehlgeschlagenen Versuchen lernt.

Ungeeignet für: Aufgaben, die keine Fehler tolerieren können, wie etwa chirurgische Eingriffe, da der Trial-and-Error-Ansatz in der Trainingsphase zu riskant wäre.

Neurosymbolic AI

Bestens geeignet für Anwendungen, die sowohl logisches Schlussfolgern als auch Lernen aus Daten erfordern, da sie symbolische Regeln mit neuronalen Netzen kombiniert. Diese Hybridarchitektur ermöglicht es, das Beste aus beiden Welten zu nutzen: die Flexibilität des maschinellen Lernens und die Präzision regelbasierter Systeme.

Beispiel: Ein Neurosymbolic AI-System kann in der Qualitätskontrolle nicht nur Defekte in Produkten erkennen (neuronale Komponente), sondern auch basierend auf vordefinierten Regeln automatisch entscheiden, ob das Produkt nachgebessert oder aussortiert werden muss (symbolische Komponente).

Ungeeignet für: Kreative Aufgaben wie Kunstgenerierung, da die symbolischen Regeln die künstlerische Freiheit zu stark einschränken würden.



Large Language Models (LLMs)

Large Language Models (LLMs) basieren auf statistischen Mustererkennungsverfahren. Sie identifizieren Korrelationen in Trainingsdaten und wenden diese auf neue Datensätze an. Als technische Grundlage dient die von Google entwickelte Deep-Learning-Architektur "Transformer" (Vaswani et al., 2017).

Das zentrale Element dieser Architektur ist der Aufmerksamkeitsmechanismus (Attention). Der Transformer wandelt zunächst Text in sogenannte Token um - diskrete numerische Einheiten, die Wörter oder Wortteile repräsentieren. Diese Token werden anschließend durch Worteinbettungen (Word Embeddings) in hochdimensionale Vektoren transformiert, welche die semantische Bedeutung und Beziehungen zwischen den Wörtern abbilden. Bei GPT-4 beispielsweise verfügen die Vektoreinbettungen über 1536 Dimensionen - eine Komplexität, die die menschliche Vorstellungskraft bei weitem übersteigt.

LLMs liefern ein statistisch plausibles nächstes Token zu einer gegebenen Eingabesequenz.

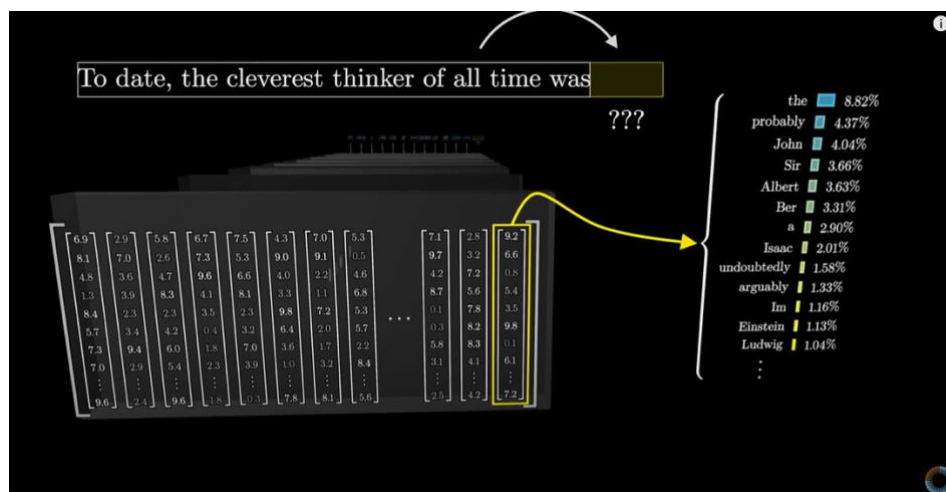


Abbildung 2: Visuelle Einführung in Transformers. [Quelle](#)

Die Plausibilität eines Tokens basiert auf den erkannten Mustern in den zugrundeliegenden Trainingsdaten. Gängige LLMs nutzen dabei nahezu sämtliche im Internet veröffentlichten Texte als Trainingsgrundlage - von hochwertigen Fachpublikationen bis hin zu qualitativ minderwertigen Beiträgen aus Online-Foren.

Durch manuelles menschliches Feedback (Reinforcement Learning from Human Feedback) wird anschließend versucht, dem LLM Leitplanken zu setzen und die Wahrscheinlichkeit unerwünschter Ausgaben zu minimieren (vgl. Perrigo, 2023).

McCoy et al. von der Princeton University formulieren es nüchtern: "...in order to develop a holistic understanding of these systems we need to consider the problem that they were trained to solve: next-word prediction over Internet text."



Die von Herstellern häufig überhöht dargestellte Leistungsfähigkeit von Large Language Models kann bei technischen Entscheidungsträgern den irreführenden Eindruck erwecken, LLMs seien universelle Problemlöser für nahezu alle Aufgabenstellungen.

Im Folgenden werden die wichtigsten Einschränkungen samt empirischer Belege dargestellt.

Zuverlässigkeit und Relevanz

Eine zentrale Einschränkung von Large Language Models liegt in ihrer fehlenden Zuverlässigkeit. Diese Systeme erkennen zwar statistische Korrelationen in Daten, ihnen fehlt jedoch das Verständnis für kausale Zusammenhänge. Sie operieren ohne Konzept von Wahrheit oder Falschheit und generieren lediglich Inhalte, die aufgrund statistischer Muster plausibel erscheinen. Diese Arbeitsweise führt zu einem charakteristischen Problem: LLMs produzieren mitunter falsche, aber überzeugend klingende Aussagen - ein Phänomen, das als "Halluzination" bezeichnet wird (Hicks et al., 2024; Dahl et al., 2024; Huang et al., 2023).

Eine weitere wesentliche Limitation betrifft die inhaltliche Relevanz der generierten Informationen. Trotz ihres beeindruckend breiten Wissensspektrums bleibt das Wissen der LLMs oberflächlich. Den Modellen fehlt ein fundamentales Verständnis der zugrundeliegenden Prinzipien und Konzepte der jeweiligen Fachdomänen. Im Unternehmenskontext ist zu beachten: LLMs verfügen über keinerlei Kenntnis spezifischer Prozesse, Strukturen oder interner Abläufe einzelner Unternehmen.

Die gängige Strategie zur Überwindung dieser Wissenslimitationen basiert auf der gezielten Anreicherung der Eingabe, dem sogenannten "Prompt". Dabei erhält das Modell alle relevanten Informationen, die es für eine fundierte Antwort benötigt. Diese als "Prompt Engineering" bekannte Technik ermöglicht es, das LLM in einen spezifischen Kontext zu setzen und sein Verhalten durch „In-Context-Learning“ zu steuern.

Um diesen Ansatz effizient und skalierbar zu gestalten, kommen zunehmend "Retrieval Augmented Generation Pipelines" (RAG) zum Einsatz. Diese automatisierten Systeme erweitern den Prompt dynamisch mit semantisch relevanten Dokumenten für die jeweilige Fragestellung. Sie schlagen damit die Brücke zwischen dem generischen Wissen des Modells und den spezifischen Informationsbedürfnissen der Anwendung.



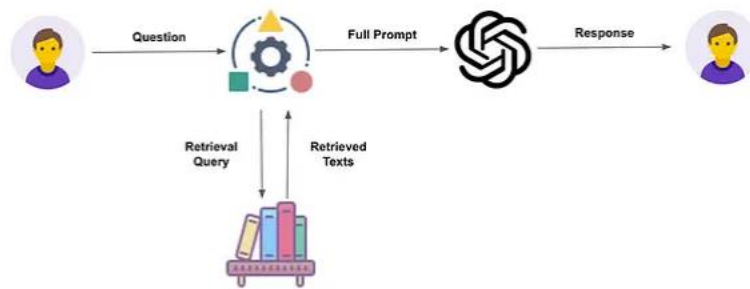


Abbildung 3: Generische Darstellung einer RAG Pipeline

In der Praxis ist die Implementierung von Retrieval Augmented Generation Pipelines mit folgenden zentralen Herausforderungen verbunden:

- ▶ Dokumentenaufbereitung: Die Segmentierung von Dokumenten in optimale Einheiten ("Chunks") unter Beibehaltung des semantischen Kontexts.
- ▶ Vektorindexierung: Die Auswahl und Konfiguration effizienter Embedding-Modelle und Ähnlichkeitssuchen für präzise Retrieval-Ergebnisse.
- ▶ Prompt Engineering: Die effektive Integration der abgerufenen Informationen in den Prompt bei gleichzeitiger Minimierung von Halluzinationen.
- ▶ Skalierbarkeit: Die Gewährleistung kurzer Antwortzeiten und effizienter Ressourcennutzung auch bei großen Dokumentenmengen.
- ▶ Qualitätssicherung: Die kontinuierliche Überwachung der Antwortqualität und Aktualität der Wissensbasis.

Zusammenfassend lässt sich festhalten: Die mangelnde Zuverlässigkeit und Relevanz des generierten Outputs stellt eine inhärente Limitation von LLMs dar. Die verfügbaren Lösungsstrategien sind in der betrieblichen Praxis mit erheblichen Implementierungshürden verbunden.

Abstraktionsvermögen und „Reasoning“

Ein zentrales Merkmal menschlicher Intelligenz ist das Abstraktionsvermögen - die Fähigkeit, aus konkreten Erfahrungen allgemeine Prinzipien abzuleiten und diese flexibel auf neue Situationen zu übertragen. Diese Form der Generalisierung ermöglicht es, erlernte Konzepte und Lösungsstrategien auf unbekannte Aufgabenstellungen anzuwenden, selbst wenn diese sich in ihrer konkreten Ausprägung von den ursprünglichen Lernerfahrungen unterscheiden.

Während aktuelle KI-Systeme in spezifischen Bereichen beeindruckende Ergebnisse liefern, fehlt ihnen dieses grundlegende Abstraktionsvermögen weitgehend. Dies kann anhand ihrer mathematischen Fähigkeiten nachvollzogen werden. Die Evaluation mathematischer Fähigkeiten von LLMs erfolgt häufig



mittels standardisierter Benchmarks wie GSM8k. Aktuelle Modelle erreichen bei diesen Tests nahezu optimale Ergebnisse.

Mirzadeh et al. (2024) haben eine kritische Überprüfung dieser Benchmarks vorgenommen. Die Studie basiert auf der Hypothese, dass ein System mit echtem Problemverständnis zwei zentrale Eigenschaften aufweisen sollte:

1. Invarianz gegenüber irrelevanten Änderungen in der Aufgabenstellung.
2. Fähigkeit zur Unterscheidung zwischen wesentlichen und unwesentlichen Informationen.

Zur Überprüfung dieser Hypothese wurden zwei Testmethoden implementiert:

Ad 1) Invarianzprüfung:

- ▶ Variation nicht-mathematisch relevanter Parameter
- ▶ Analyse der Stabilität der Lösungsfindung bei oberflächlichen Änderungen

Ad 2) Kontextrelevanz-Analyse:

- ▶ Integration scheinbar relevanter, aber lösungsirrelevanter Zusatzinformationen
- ▶ Evaluation der Fähigkeit zur Differenzierung zwischen wesentlichen und unwesentlichen Informationen

Diese methodische Herangehensweise ermöglicht eine systematische Unterscheidung zwischen:

- ▶ Echtem mathematischem Verständnis
- ▶ Kontextunabhängiger Problemlösungskompetenz
- ▶ Oberflächlicher Mustererkennung

Im Ergebnis zeigen LLMs deutliche Leistungsschwankungen bei gleichen Fragen mit unterschiedlichen Zahlenwerten. Weiters verschlechtern sich die Antworten, je nachdem wie die Aufgabe formuliert ist.

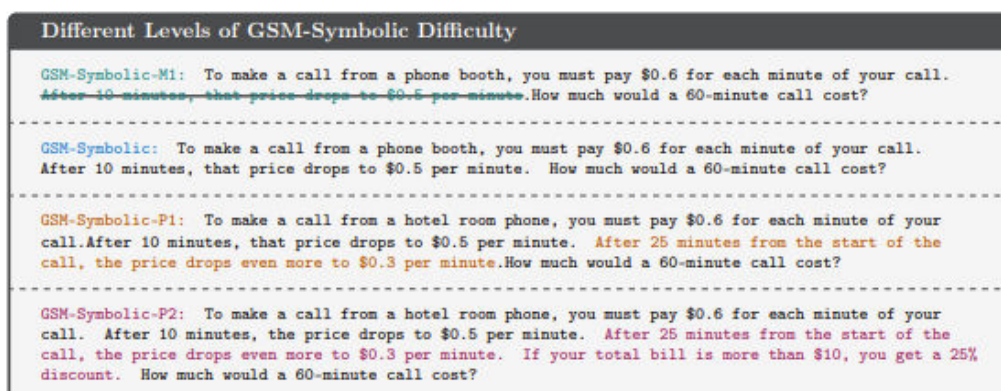


Abbildung 4: Je länger die Anweisung wird, desto schlechter ist die Performance des LLMs. [Quelle](#)



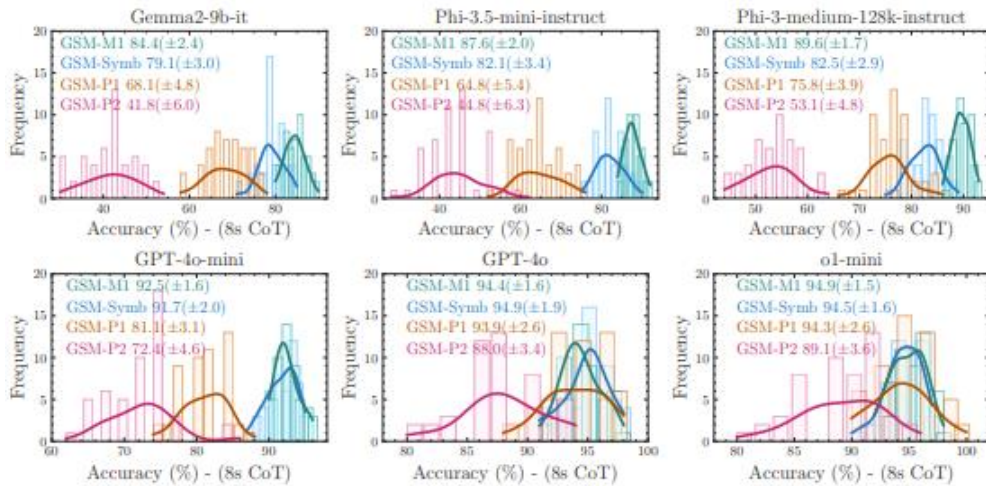


Abbildung 5: Performance basierend auf Input aus Abbildung 4. Quelle

GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

o1-mini: Oliver picks kiwis over three days as follows:
 - Friday: 44 kiwis
 - Saturday: 58 kiwis
 - Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis
 However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Llama3-8B: Let's break it down step by step:
 Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is $2 \times 44 = 88$ kiwis.
 Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: $88 - 5 = 83$ kiwis. Now, let's add up the total number of kiwis Oliver has:
 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis
 So, Oliver has 185 kiwis in total.

Abbildung 6: Zusätze in der Aufgabenstellung führen zu einem Leistungsabfall von bis zu 65%. [Quelle](#)

Das Hinzufügen selbst eines einzelnen, für die Lösung irrelevanten Details führt zu einer drastischen Verschlechterung der Modellleistung um bis zu 65% (siehe Abbildung 6). Diese Anfälligkeit deutet darauf hin, dass die Modelle keine echte konzeptionelle Logik entwickeln, sondern statistische Muster aus ihren Trainingsdaten reproduzieren (Mirzadeh et al., 2024; Jiang et al., 2024).

Zusammenfassend ist für technische Entscheidungsträger in Unternehmen zentral zu verstehen: Large Language Models (LLMs) stellen eine wichtige Lösungskomponente dar, nicht jedoch die komplette Lösung selbst. Sie sind keine Universallösung, sondern erfordern eine sorgfältige technische, inhaltliche und organisatorische Einbettung in bestehende Systeme und Prozesse.



Transparenz

Die Harvard Data Science Review veröffentlichte Ende Februar 2024 in Kooperation mit Microsoft Research ihr Special Issue 5 zum Thema Transparenz von Large Language Models (LLMs). Die zentrale Erkenntnis lautet: While capturing intense public and academic enthusiasm, large language models (LLMs) suffer from a lack of transparency. We still do not have clear answers to even basic questions: What can an LLM do? How well can it do it? How exactly does it work internally? This poses challenges for different stakeholders. For instance, developers of LLMs and LLM-infused applications have difficulties debugging and determining responsible use of these models; end-users often lack a sufficient understanding of these models to interact effectively and know how much to trust them; and policymakers require more transparency to have effective oversight of LLMs” (Liao et al., 2024).

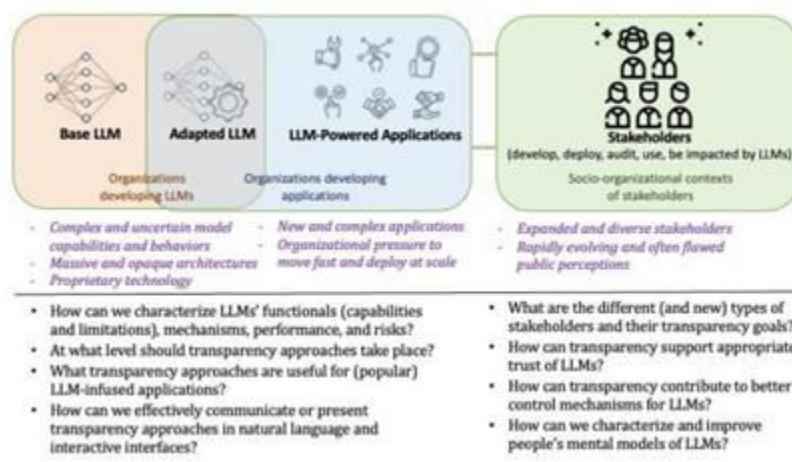


Abbildung 7: Zusammenfassung der allgemeinen offenen Fragen laut Microsoft Research. [Quelle](#)

Obwohl Large Language Models (LLMs) beeindruckende Ergebnisse liefern, bleibt ihre innere Funktionsweise weitgehend ein Rätsel. Zwar gibt es Forschungsbemühungen, insbesondere von Anthropic, die Wissensrepräsentation in LLMs zu entschlüsseln (Templeton et al., 2024), jedoch sind diese Erkenntnisse für Unternehmen in der praktischen Anwendung kaum nutzbar. Weiters ist dabei zu beachten, dass beobachtete Korrelationen nicht automatisch kausale Zusammenhänge implizieren.

Die mangelnde Transparenz bei der Entscheidungsfindung von LLMs birgt das Risiko, dass Vorurteile, Fehler oder Mängel in ihrer Urteilsfindung nicht entdeckt werden und unkontrolliert bleiben. Ohne Einblick in ihre Funktionsweise ist die Möglichkeit für Debugging und Wartung auf Trial and Error beschränkt. Dies stellt erhebliche Risiken dar, wenn LLMs in sensiblen Anwendungsbereichen eingesetzt werden. Als Black Boxes lassen sich ihre Entscheidungen nicht auf Stichhaltigkeit und ethische Vertretbarkeit überprüfen. Folglich könnten fehlerhafte, unangemessene oder ethisch bedenkliche Entscheidungen getroffen werden, ohne dass dies rechtzeitig erkannt wird.



Planungsfähigkeit

LLMs können kaum oder gar nicht planen. Aktuelle Forschungsergebnisse dokumentieren, dass die Leistungsfähigkeit der Modelle bei zunehmender Komplexität der Planungsaufgaben signifikant abnimmt. Dies gilt auch für fortgeschrittene Systeme wie o1-preview von OpenAI. (Vgl. Valmeekam et al., 2024, 1).

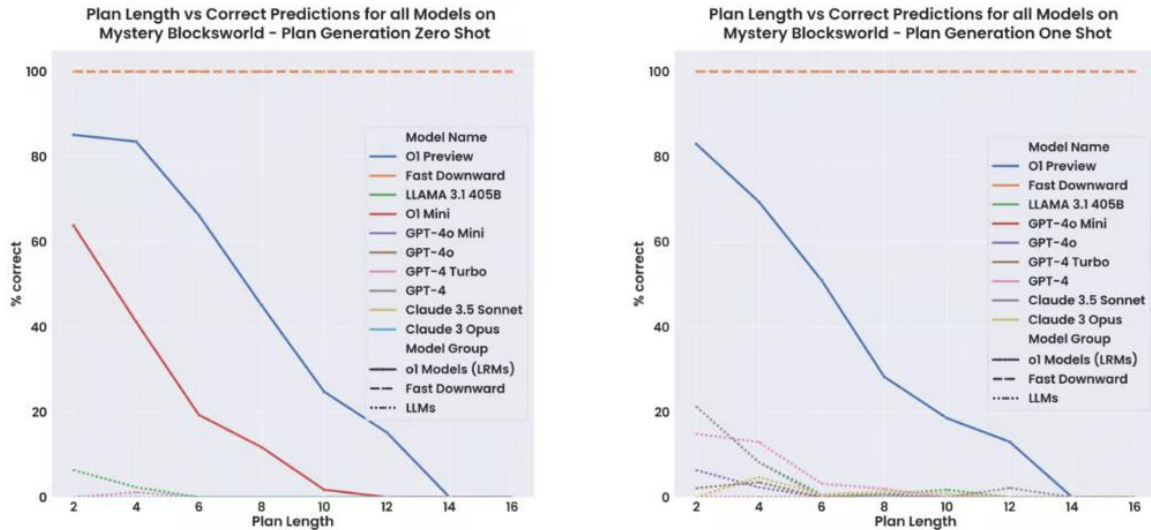


Abbildung 8: Die Planungsfähigkeit von LLMs nimmt mit der Planlänge ab. [Quelle](#)

Eine vergleichende Analyse zwischen modernen LLMs und klassischen Planungsalgorithmen zeigt signifikante Leistungsunterschiede. Der „Fast Downward“ Algorithmus aus 2006, der auf deterministischer Suche basiert, erreicht in Benchmark-Tests eine Erfolgsquote von 100 Prozent. Im Vergleich dazu erzielt das LLM o1-preview in komplexen Planungsszenarien eine Erfolgsquote von 37,3 Prozent.

| Domain | Shots | Instances correct | | | Average Time Taken (in secs) | | |
|---------------------|-----------|--------------------|--------------------|-------------------|------------------------------|---------|---------------|
| | | o1-preview | o1-mini | Fast Downward | o1-preview | o1-mini | Fast Downward |
| Blocksworld | Zero Shot | 587/600 (97.8%) | 340/600 (56.6%) | 600/600 (100%) | 40.43 | 10.84 | 0.265 |
| | One Shot | 247/600 (41.6%) | - | 600/600 (100%) | 82.03 | - | 0.265 |
| Mystery Blocksworld | Zero Shot | 317/600 (52.8%) | 115/600 (19.1%) | 600/600 (100%) | 83.37 | 35.54 | 0.265 |
| | Zero Shot | 224/600 (37.3%) | - | 600/600 (100%) | 111.11 | - | 0.265 |

Abbildung 9: Leistungsfähigkeit von LLMs bei der Lösung von „Blocksworld“ Tasks. [Quelle](#)



Abschließend lässt sich zur Planungsfähigkeit von Large Language Models (LLMs) auf Basis aktueller Erkenntnisse festhalten: Sie verfügen über kaum nennenswerte Planungsfähigkeiten und erfüllen auch diese zentrale Anforderung an intelligente Systeme (noch) nicht. Ihre scheinbare Planungskompetenz bleibt lediglich oberflächlich.

Vertrautheit vs. Komplexität

Das Verhältnis zwischen Aufgabenvertrautheit und Komplexität bei der Bewertung von KI-Systemen wirft grundlegende Fragen auf, wie der Google KI-Forscher François Chollet darlegt. Seine Kernthese: Nicht die Komplexität einer Aufgabe, sondern die Vertrautheit damit sollte als Maßstab für die Leistungsfähigkeit von LLMs herangezogen werden.

Diese These lässt sich anhand eines anschaulichen Beispiels verdeutlichen: Ein 12-jähriges Kind kann durch reines Auswendiglernen durchaus in der Lage sein, mathematische Probleme auf Doktoranden-Niveau zu lösen - vorausgesetzt, die Aufgabenstellung bleibt weitestgehend gleich. Wird dasselbe Kind jedoch mit einer unbekannteren, wenn auch einfacheren Mathematikaufgabe auf Bachelor-Niveau konfrontiert, scheitert es. Der Grund liegt darin, dass Auswendiglernen zwar die Bewältigung komplexer, bekannter Aufgaben ermöglicht, jedoch kein echtes Verständnis voraussetzt.

Die gegenwärtig verwendeten Benchmarks konzentrieren sich überwiegend auf die Ergebnisse und vernachlässigen dabei den Weg zur Lösung. Dies führt zu einer systematischen Überschätzung der tatsächlichen "Intelligenz" von LLMs. Denn ähnlich wie das Kind können diese Systeme durch extensives Training komplexe Aufgabenmuster erkennen und reproduzieren, ohne ein tieferes Verständnis zu entwickeln.



| Name | Score (public eval) | Verification Score (semi-private eval) |
|-----------------|---------------------|--|
| Ryan Greenblatt | 42% | 43% |
| icecuber 2020 | 39% | 17% |
| o1-preview | 21% | 16% |
| Claude 3.5 | 21% | 14% |
| o1-mini | 13% | 9.5% |
| GPT-4o | 9% | 5% |
| Gemini 1.5 | 8% | 4.5% |

Abbildung 10: Das Leaderboard der ARC-AGI Benchmark nach Francois Chollet am 23.10.2024



Um die wahre Intelligenz eines KI-Systems zu ermitteln, muss es daher in Situationen getestet werden, in denen es nicht auf bekannte Muster zurückgreifen kann. Nur wenn ein System unter diesen Bedingungen erfolgreich ist, kann von einer "intelligenten" Verarbeitung gesprochen werden. Der von Chollet entwickelte ARC-AGI (Abstract Reasoning Corpus-Artificial General Intelligence) Benchmark verfolgt genau diesen Ansatz. Die Ergebnisse sind ernüchternd: Selbst die fortschrittlichsten LLMs wie o1-preview erreichen hier lediglich eine Erfolgsquote von 21%.



Conclusio

Aktuelle Large Language Models (LLMs) zeigen beeindruckendes "Pattern-Matching" auf hohem Niveau, entwickeln jedoch bis dato keine Intelligenz. Diese Einschätzung basiert auf der Analyse ihrer Zuverlässigkeit, Relevanz, ihres Abstraktionsvermögens sowie ihrer Reasoning- und Planungsfähigkeiten.

LLMs entfalten ihr volles Potenzial in den Anwendungsbereichen, für die ihre Architektur konzipiert und ihr Training optimiert wurde - insbesondere bei der Textgenerierung, Zusammenfassung, Übersetzung u.ä.m. Bei der Automatisierung komplexer Geschäftsprozesse stoßen sie hingegen an deutliche Grenzen.

Für technische Entscheidungsträger ergeben sich daraus folgende Handlungsempfehlungen:

- ▶ Setzen Sie KI-Technologien gezielt und anwendungsspezifisch ein
- ▶ Verstehen Sie LLMs als Baustein einer Gesamtlösung, nicht als Komplettlösung
- ▶ Entwickeln Sie flexible, robuste Systemarchitekturen
- ▶ Implementieren Sie eine Microservice-Architektur mit austauschbaren Komponenten, um der rasanten technologischen Entwicklung im KI-Bereich gerecht zu werden



Quellen

Chollet, F. (2019). On the Measure of Intelligence. Google

Dahl, M. et al. (2024). Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive. Stanford University. <https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive>

Hicks, M. et al. (2024). ChatGPT is bullshit. Ethics and Information Technology.

Huan, L. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. Harbin Institute of Technology

Jaffri, A. et al. (2024). Hype Cycle for Artificial Intelligence, 2024. Gartner

Jiang, B. (2024). A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners. University of Pennsylvania

Liao, Q. et al. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. Harvard Data Science Review

McCoy, R. (2023). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. Princeton University

Mirzadeh, I. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. Apple

Mullen, A. (2024). Artificial Intelligence Primer for 2024. Gartner.

Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic. Time. <https://time.com/6247678/openai-chatgpt-kenya-workers/>

Ramos, L. et al. (2024). When Not to Use Generative AI. Gartner.

Templeton, A. et al. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Anthropic https://transformer-circuits.pub/2024/scaling-monosemanticity/?utm_source=thetechoasis.beehiiv.com&utm_medium=newsletter&utm_campaign=the-4-jobs-of-the-future&_bhlid=bf66a8164acdd798329c7773519b279afcbffcd4/

Valmeekam, K. (2024). LLMS STILL CAN'T PLAN; CAN LRMS? A PRELIMINARY EVALUATION OF OPENAI'S O1 ON PLANBENCH. SCAI, Arizona State University

Vaswani, A. et al. (2017). Attention Is All You Need. Google.



Abbildungsverzeichnis

| | |
|---|----|
| Abbildung 1: GenAI ist ein Teilbereich von KI | 4 |
| Abbildung 2: Visuelle Einführung in Transformers. Quelle | 7 |
| Abbildung 3: Generische Darstellung einer RAG Pipeline | 9 |
| Abbildung 4: Je länger die Anweisung wird, desto schlechter ist die Performance des LLMs. Quelle | 10 |
| Abbildung 5: Performance basierend auf Input aus Abbildung 4. Quelle | 11 |
| Abbildung 6: Zusätze in der Aufgabenstellung führen zu einem Leistungsabfall von bis zu 65%. Quelle | 11 |
| Abbildung 7: Zusammenfassung der allgemeinen offenen Fragen laut Microsoft Research. Quelle | 12 |
| Abbildung 8: Die Planungsfähigkeit von LLMs nimmt mit der Planlänge ab. Quelle | 13 |
| Abbildung 9: Leistungsfähigkeit von LLMs bei der Lösung von „Blocksworld“ Tasks. Quelle | 13 |
| Abbildung 10: Das Leaderboard der ARC-AGI Bechhmark nach Francois Chollet am 23.10.2024 | 14 |

