P/A Media

Das Geheimnis hinter LLM-Halluzinationen

und wie man sie bekämpft

Thomas Reißer
Head of SEO bei PIA Media GmbH in München

Juli 2024

inhalt

Was genau sind Large Language Models?

Large Language Models (LLMs) stellen eine faszinierende Entwicklung im Bereich der künstlichen Intelligenz dar. Diese Modelle durchlaufen ein aufwendiges Training mit riesigen Datensätzen und bestehen aus komplexen neuronalen Netzen in vielen Schichten, die es ermöglichen, Muster in den Daten zu erkennen und auf neue Ein-gaben anzuwenden. Eine ihrer bemerkenswertesten Fähigkeiten ist die Generierung verschiedener kreativer Textformate, darunter Gedichte, Code, Skripte, E-Mails und Briefe. Darüber hinaus sind LLMs in der Lage, Fragen umfassend und informativ zu beantworten, selbst wenn sie offen, komplex oder ungewöhnlich sind. Eine weitere bedeutende Fähigkeit besteht darin, Sprachen zu übersetzen, wobei sie sowohl den semantischen Inhalt als auch den Stil und Ton der Originalsprache erhalten.

Dennoch ist es wichtig zu betonen, dass LLMs keinen echten kreativen Prozess durchführen. Sie generieren Text auf der Grundlage von Wahrscheinlichkeiten und der Anzahl der Parameter im Training und verwenden ausschließlich Inhalte, die in den Trainingsdaten enthalten waren. Während sie beeindruckende Ergebnisse liefern und eine breite Palette von Aufgaben bewältigen können, beruht ihre Leistung auf dem effizienten Mustererkennungssystem, das sie während des Trainings entwickeln. Die Tatsache, dass sie keine neuen Ideen oder Konzepte hervorbringen, sondern lediglich vorhandene Informationen neu kombinieren, ist ein wichtiger Aspekt, der bei der Beurteilung ihrer Fähigkeiten berücksichtigt werden sollte.

Was sind LLM-Halluzinationen?

Ein wichtiger Aspekt bei der Analyse von LLMs sind die sogenannten LLM-Halluzinationen, die auftreten, wenn ein generierter Text zwar syntaktisch korrekt, flüssig und natürlich erscheint, jedoch faktisch falsch, unsinnig oder unzutreffend für die gegebene Eingabe ist.

LLM-Halluzinationen unterscheiden sich von anderen Fehlern durch ihre intrinsische Falschheit und Irreführung. Sie beinhalten die Generierung von Informationen, die nicht in der Realität verankert sind, unabhängig von der Richtigkeit der zugrundeliegenden Daten.

Diese Unterscheidung ist wichtig, um das Potenzial und die Grenzen von LLMs genau zu verstehen und kritisch zu bewerten. In vielen Anwendungsfällen, insbesondere solchen, die auf präzisen und korrekten Informationen beruhen, wie etwa medizinische oder rechtliche Beratung, können LLM-Halluzinationen zu erheblichen Problemen führen. Daher ist es von entscheidender Bedeutung, Mechanismen zu entwickeln, um solche Fehler zu erkennen und zu minimieren, um das Vertrauen in die Zuverlässigkeit von LLMs zu stärken.

Warum ist es wichtig, das Problem anzugehen?

Halluzinationen von LLMs können schwerwiegende Folgen haben, wie die Verbreitung von Fehlinformationen, Datenschutzverletzungen und Sicherheitsbedenken für reale Anwendungen. Beispielsweise kann ein aus Patienteninformationen generierter, halluzinierter Bericht im medizinischen Bereich ein ernstes Risiko für den Patienten darstellen. Derartige Halluzinationen beeinträchtigen letztlich auch das generelle Vertrauen der User in diese Technik, weshalb es wichtig ist, dieses Problem schnell anzugehen.

Ein reales Beispiel ist der Fall eines Professors der Texas A&M University, der seine gesamten Studierenden durchfallen ließ, nachdem ChatGPT fälschlicherweise behauptete, ihre Arbeiten seien von der KI geschrieben worden. Dies führte dazu, dass vielen Studierenden ihre Diplome verweigert wurden, was nicht nur ihre akademische Laufbahn gefährdete, sondern auch das Vertrauen in die Zuverlässigkeit solcher Technologien untergrub. [1]

Ebenso kam es durch ChatGPT zu einer falschen Anschuldigung wegen sexueller Belästigung gegen den Juraprofessor Jonathan Turley von der George Washington University. Das KI-Modell erfand einen nicht existierenden Artikel der Washington Post und beschuldigte Turley fälschlicherweise, eine Studentin während einer Klassenreise belästigt zu haben. Solche Vorfälle zeigen, wie gefährlich und irreführend KI-generierte Inhalte sein können, und betonen die Notwendigkeit, strengere Überprüfungsmechanismen und ethische Richtlinien für den Einsatz von KI zu etablieren. [2]

Welche Ursachen führen zu Halluzinationen?

Die Qualität der Trainingsdaten spielt eine entscheidende Rolle für die Leistung von Large Language Models (LLMs). Inkonsistenzen, Ungenauigkeiten, Voreingenommenheit und Vorurteile in den Trainingsdaten können erhebliche Auswirkungen haben, da sie zu Fehlern in der Generierung führen können. Wenn das Modell falsche Informationen aus den Trainingsdaten übernimmt, kann dies zu ungenauen oder irreführenden Ausgaben führen. Ebenso können unvollständige Datensätze, die eine begrenzte Diversität und Kontextinformationen aufweisen, die Fähigkeit des LLMs einschränken, relevante und präzise Ergebnisse zu erzielen.

Während es wichtig ist, die Diversität in den Trainingsdaten zu fördern, um die Leistung des Modells zu verbessern, kann eine übermäßige Korrektur zur Diversität auch zu falschen Ergebnissen führen, da möglicherweise zu viele unzusammenhängende oder nicht repräsentative Datenpunkte in das Training einbezogen werden. Daher ist eine sorgfältige Auswahl und Bereinigung der Trainingsdaten von entscheidender Bedeutung, um sicherzustellen, dass das LLM optimale Ergebnisse liefert.

1. Eingabe: Vage oder unpräzise Formulierungen sowie ein Mangel an Kontext in der Eingabe können dazu führen, dass das Large Language Model (LLM) die

Absicht des Benutzers missversteht. Dies wiederum kann zu unvorhersehbaren oder fehlerhaften Ergebnissen bei der Generierung von Text führen. Es ist daher wichtig, klare und präzise Eingaben zu formulieren und zusätzlichen Kontext bereitzustellen, um die Genauigkeit und Relevanz der vom LLM generierten Ausgaben zu verbessern.

Beispiel Rechtliche Beratung: „Ist es in Ordnung, meine Nachbarn zu filmen?“ Die Antwort könnte variieren, je nachdem, ob die Person in einem öffentlichen Raum oder in ihrem eigenen Haus filmt, ob es um Überwachung geht oder um persönliche Aufnahmen. Ein AI-Modell könnte ohne genügend Kontext Halluzinationen erzeugen und Annahmen darüber treffen, was rechtlich akzeptabel ist. Dies könnte zu falschen Schlussfolgerungen führen und den Benutzer möglicherweise in rechtliche Schwierigkeiten bringen.

2. Kontext: Bei längeren Gesprächen besteht die Gefahr, dass das Large Language Model (LLM) den Zusammenhang verliert und widersprüchliche Aussagen generiert. Zudem kann das Modell Schwierigkeiten haben, den Kontext angemessen zu berücksichtigen und die Generierung an die spezifische Situation anzupassen. Dies kann zu unpassenden oder nicht relevanten Antworten führen, da der Kontext entscheidend ist, um die Intention des Nutzers zu verstehen.

3. Faktenwissen: Obwohl Large Language Models (LLMs) auf umfangreiche Datenmengen zugreifen können, ist ihre Informationsbasis dennoch begrenzt. Sie verfügen zwar über ein beachtliches Wissensrepertoire, doch sie haben keinen Zugriff auf das gesamte Wissen der Welt. Dies kann dazu führen, dass LLMs bei bestimmten Themen oder spezialisierten Bereichen unvollständige oder fehlerhafte Informationen liefern. Darüber hinaus können LLMs Schwierigkeiten bei der Verifizierung von Fakten haben, da falsche Informationen möglicherweise im Nachhinein schwer zu erkennen sind. Dies unterstreicht die Notwendigkeit einer kritischen Prüfung der von LLMs generierten Inhalte und betont die Bedeutung von menschlicher Überwachung und Validierung, insbesondere bei sensiblen oder entscheidenden Informationen.

Beispiel: Angenommen, ein Nutzer fragt ein Large Language Model nach dem Gewicht des Mount Everest. Das Modell antwortet daraufhin mit einer präzisen Zahl, die scheinbar korrekt ist. Doch bei genauerer Überprüfung stellt sich heraus, dass das Modell lediglich auf eine falsche Information aus seinen Trainingsdaten zurückgegriffen hat. Tatsächlich ändert sich das Gewicht des Mount Everest im Laufe der Zeit aufgrund von geologischen Veränderungen, sodass eine statische Zahl keine genaue Antwort darstellt. In diesem Beispiel zeigt sich deutlich die Schwierigkeit von LLMs, komplexe und sich verändernde Fakten korrekt zu erfassen und darauf basierend genaue Antworten zu liefern.

4. „Lost-in-the-Middle“[3]: In der Praxis zeigt sich, dass selbst Large Language Models mit großen Kontextfenstern von 8K bis 32K Token relevante Details im Text, insbesondere in der Mitte, auslassen können.

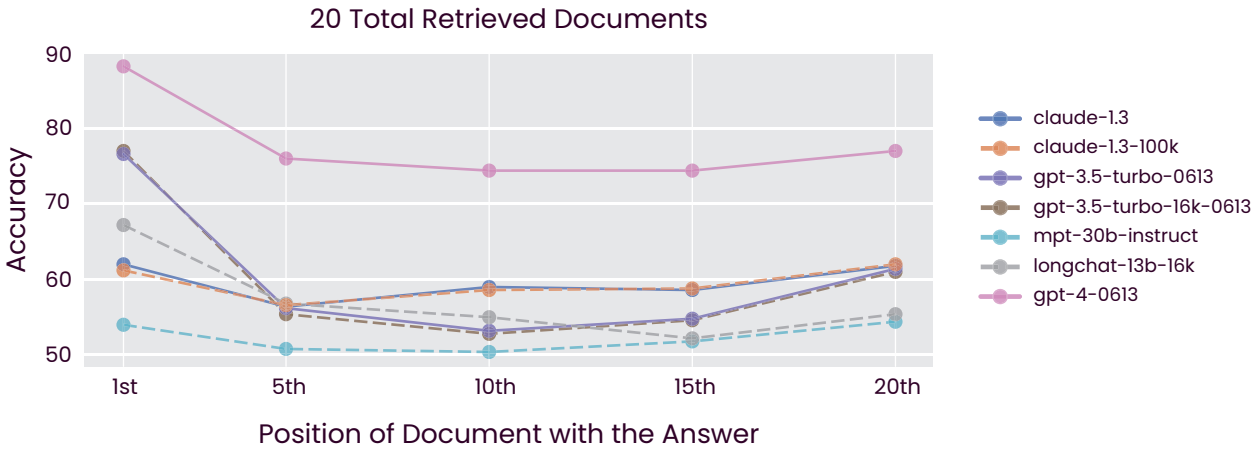

Abbildung 1: Lost-in-the-Middle Analyse (https://arxiv.org/pdf/2307.03172)

Dies kann zu unvollständigen oder ungenauen Ergebnissen oder Fehlinterpretationen führen. Die Ursache liegt in der begrenzten Verarbeitungskapazität des Modells.

Ein Kontextfenster von 8K-32K Token mag zwar groß klingen, aber es reicht oft nicht aus, um alle relevanten Informationen in einem längeren Text zu erfassen. Daher kann es passieren, dass der LLM den Zusammenhang verliert oder wichtige Details übergeht, was die Qualität der generierten Ergebnisse beeinträchtigen kann. Dieser Aspekt verdeutlicht die Herausforderungen, denen LLMs gegenüberstehen, und die Notwendigkeit, ihre Fähigkeiten zur Berücksichtigung von Kontexten kontinuierlich zu verbessern, um präzisere und umfassendere Antworten zu erzielen.

5. „Reasoning Gap“: Die „Denklücke“ bei aktuellen Large Language Models (LLMs) offenbart eine Diskrepanz zwischen ihrer Leistungsfähigkeit bei bekannten im Vergleich zu spontan zu lösenden Problemen.

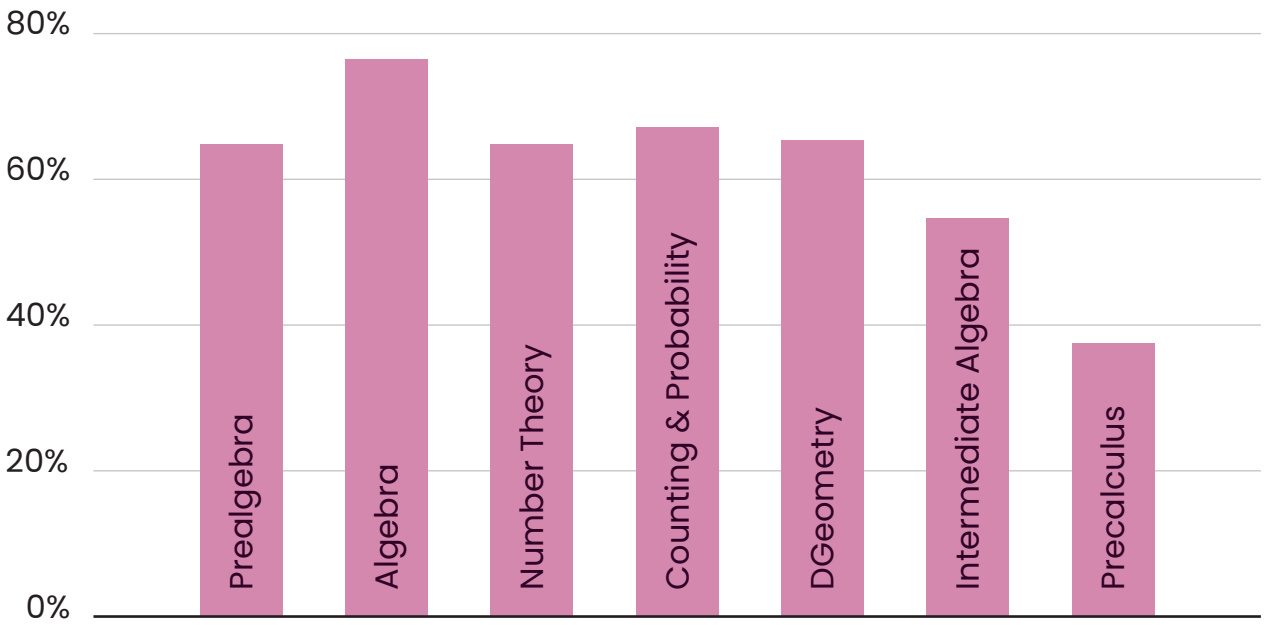

Abbildung 2: Gap across subjects, aggregated across all models.
(https://arxiv.org/pdf/2402.19450)

Diese Diskrepanz wird deutlich, wenn man sich die Leistung von LLMs wie GPT-4 betrachtet, welches nur etwa 10% der 5.000 dynamischen Aufgaben erfolgreich löst. Aktuelle LLMs befinden sich in der ersten von vier Stufen der Generalisierungsfähigkeit.

Sie können Antworten auf eine statische Menge von Aufgaben geben und interpolieren dazwischen, jedoch fehlt ihnen ein echtes Verständnis für die Probleme. Dies verdeutlicht die bestehenden Grenzen in der Fähigkeit von LLMs, spontane oder dynamische Probleme zu lösen, und unterstreicht die Notwendigkeit weiterer Forschung und Entwicklung, um die kognitive Fähigkeit und das Problemlösungsvermögen dieser Modelle zu verbessern.[4]

6. KI-Anfälligkeit gegen „Silent Data Corruptions“[5] (SDCs): KI-Modelle sind anfällig gegenüber Fehlern in ihren Parametern. Diese Fehler, wie Bit-Kipp-punkte, können bei Hardwareproblemen auftreten und zu falschen Modell-vorhersagen führen. Der Parameter Vulnerability Factor (PVF) quantifiziert die Wahrscheinlichkeit, dass solche Fehler zu falschen Outputs führen. Die Ergeb-nisse zeigen, dass verschiedene Modellteile unterschiedlich anfällig sind, was für die Entwicklung robuster KI-Hardware wichtig ist.

Fehlerhafte SDCs während des Trainings von Large Language Models (LLMs) können die Genauigkeit und Zuverlässigkeit dieser Modelle erheblich beein-trächtigen. SDCs sind besonders problematisch, da sie oft unbemerkt bleiben und somit schwer zu diagnostizieren sind. Wenn Parameter in LLMs durch SDCs korrumpiert werden, können die Modelle ungenaue oder falsche Ausgaben erzeugen. Daher ist es wichtig, Mechanismen zur Erkennung und Korrektur sol-cher Fehler zu entwickeln, um die Integrität der Modelle zu gewährleisten.

7. „Jailbreak“-Techniken: „Jailbreak“-Techniken bezeichnen externe Manipu-lationen von Large Language Models (LLMs), die darauf abzielen, das System zur Generierung von illegalen Inhalten zu bringen. Diese Techniken können ver-schiedene Methoden umfassen, von der gezielten Einführung von fehlerhaften Trainingsdaten bis hin zur Ausnutzung von Schwachstellen im Modell selbst.

Durch solche Manipulationen können LLMs dazu gebracht werden, unange-messene oder rechtswidrige Inhalte zu generieren, die beispielsweise Hassre-de, Pornografie oder andere gesetzwidrige Materialien enthalten. Der Einsatz von „Jailbreak“-Techniken stellt eine ernsthafte Bedrohung dar und erfordert

geeignete Sicherheitsmaßnahmen, um die Integrität und Vertrauenswürdigkeit von LLMs zu gewährleisten.

8. „ArtPrompt"-Technik: Die „ArtPrompt"-Technik bezeichnet eine Methode externer Manipulation von Large Language Models (LLMs), die darauf abzielt, das Modell dazu zu bringen, bestimmte Begriffe zu generieren, die normalerweise gesperrt oder eingeschränkt sind. Diese Technik funktioniert, indem zunächst der gesperrte Begriff durch eine Maske ersetzt wird. Im zweiten Schritt erfolgt dann die Ersetzung dieser Maske durch ASCII Art, die den gesperrten Begriff visuell repräsentiert.

Auf diese Weise können LLMs dazu gebracht werden, Inhalte zu generieren, die anstößig, beleidigend oder gesetzwidrig sind, ohne dass der gesperrte Begriff direkt verwendet wird. Die „ArtPrompt"-Technik stellt eine Herausforderung für die Sicherheit und Integrität von LLMs dar und erfordert geeignete Gegenmaßnahmen, um Missbrauch zu verhindern und die Qualität der generierten Inhalte zu gewährleisten.

Ansätze zur Minimierung von Halluzinationen

Strategien auf Betreiber-Seite

1. Datenbereinigung: Die Datenbereinigung ist ein entscheidender Schritt im Vorbereitungsprozess von Trainingsdaten für Large Language Models (LLMs). Hierbei werden inkonsistente, ungenaue, voreingenommene und irrelevante Informationen aus dem Trainingsset entfernt. Durch diesen Prozess wird sichergestellt, dass das Modell auf qualitativ hochwertige und verlässliche Daten zugreift, um genaue und konsistente Ergebnisse zu erzielen. Inkonsistenzen oder Fehler in den Trainingsdaten könnten dazu führen, dass das LLM falsche Informationen übernimmt oder ungenaue Antworten generiert, was Leistung und Vertrauenswürdigkeit des Modells beeinträchtigen würde.

2. Datenerweiterung: Die Datenerweiterung ist ein weiterer wichtiger Schritt, um die Qualität und Vielfalt der Trainingsdaten zu verbessern. Durch Hinzufügen von vielfältigen und kontextbezogenen Daten können Lücken im Trainingsset geschlossen und die Leistung des LLMs verbessert werden. Dies kann erreicht werden, indem verschiedene Quellen und Datenquellen einbezogen werden, um ein breiteres Spektrum von Informationen abzudecken und sicherzustellen, dass das Modell über ein umfassendes Wissensrepertoire verfügt. Eine sorgfältige Datenerweiterung trägt dazu bei, die Generalisierungsfähigkeit des LLMs zu verbessern und seine Fähigkeit zu stärken, verschiedene Aufgaben und Anwendungsfälle erfolgreich zu bewältigen.

In dem Paper „Agent Planning with World Knowledge Model“[7] wird im Detail auf die Möglichkeit eingegangen, wie eine gezielte Integration von relevantem „Weltwissen“ bereits während des Trainings und später der Generierung aussehen könnte.

3. Validierung und Überprüfung: Die Validierung und Überprüfung der Trainingsdaten sind essenzielle Schritte, um sicherzustellen, dass die Daten von hoher Qualität und Richtigkeit sind. Durch die Einführung von Mechanismen zur Verifizierung der Richtigkeit der Daten können potenzielle Fehler oder Inkonsistenzen frühzeitig erkannt und behoben werden. Dies kann beispielsweise durch manuelle Überprüfung, automatisierte Tests oder die Nutzung von externen Validierungsquellen erfolgen. Eine gründliche Validierung und Überprüfung gewährleistet, dass das LLM auf verlässliche und korrekte Informationen zugreift, was wiederum die Genauigkeit und Zuverlässigkeit der generierten Ergebnisse erhöht und das Vertrauen in das Modell stärkt. Einen möglichen Lösungsansatz bietet hier das Open Source Projekt FineWeb.[8] Wichtige Grundlagen Forschung für ein besseres Verständnis der inneren Abläufe von LLMs lieferte jüngst das Team bei Anthropic mit ihrer Extraktion von Millionen von Features ihres Models Claude 3 Sonnet.[9]

4. Dialog-Tracking: Beim Dialog-Tracking wird der Gesprächsverlauf gespeichert und verfolgt. Diese Technik ermöglicht es dem Large Language Model

(LLM), den Kontext eines laufenden Gesprächs besser zu verstehen und angemessen darauf zu reagieren. Indem vergangene Interaktionen berücksichtigt werden, kann das LLM relevantere und kohärentere Antworten generieren, was zu einer verbesserten Benutzererfahrung führt.[10]

a. Kontextbezogene Generierung: Die kontextbezogene Generierung bezieht sich darauf, dass das LLM seine Antwort an den aktuellen Gesprächskontext anpasst. Indem der Kontext des laufenden Gesprächs berücksichtigt wird, kann das Modell präzisere und relevantere Antworten liefern. Dadurch werden die Chancen auf Missverständnisse oder inkonsistente Antworten verringert und die Qualität der Kommunikation verbessert.

b. Erkennung von Widersprüchen: Die Erkennung von Widersprüchen ist ein wichtiger Aspekt der Dialogverarbeitung. Das LLM überprüft kontinuierlich den Gesprächsverlauf, um inkonsistente Aussagen oder widersprüchliche Informationen zu identifizieren und zu vermeiden. Durch die Vermeidung von Widersprüchen kann das Modell glaubwürdige und konsistente Antworten liefern, was die Effektivität und Verlässlichkeit der Kommunikation erhöht.

5. Manipulativen Eingabeaufforderungen: Die Erkennung von manipulativen Eingabeaufforderungen beinhaltet die Filterung und Blockierung von „Jailbreak"-Techniken, die darauf abzielen, das Large Language Model (LLM) zu unangemessenem Verhalten zu verleiten. Durch den Einsatz von Algorithmen und Sicherheitsmaßnahmen werden solche Manipulationsversuche erkannt und unterbunden, um die Integrität des Modells zu schützen und die Generierung von unangemessenen oder rechtswidrigen Inhalten zu verhindern.

6. Zugriffssteuerung: Die Begrenzung der Zugriffsmöglichkeiten betrifft die Kontrolle der Benutzer, die das LLM beeinflussen können. Durch die Implementierung von Zugriffskontrollen und Berechtigungen wird sichergestellt, dass nur autorisierte Benutzer Zugriff auf das Modell haben und es verwenden können. Auf diese Weise wird das Risiko von Missbrauch oder unerwünschter Manipulation minimiert und die Sicherheit des Systems gewährleistet.

7. Transparenz und Aufklärung: Die Transparenz und Aufklärung der Benutzer über die Funktionsweise von LLMs und die Risiken von Halluzinationen ist von entscheidender Bedeutung. Benutzer sollten über die potenziellen Risiken informiert werden, die mit der Verwendung von LLMs verbunden sind, insbesondere im Hinblick auf die Möglichkeit von Halluzinationen und die Verbreitung von irreführenden oder ungenauen Informationen. Durch eine umfassende Aufklärung können Benutzer besser verstehen, wie sie LLMs sicher und verantwortungsbewusst nutzen können, und angemessene Vorkehrungen treffen, um die Risiken zu minimieren.

Strategien auf Benutzer-Seite

Die Möglichkeit zur Optimierung des Prompts durch Large Language Models (LLMs) ist ein zentraler Aspekt, der die Leistungsfähigkeit und Vielseitigkeit dieser Modelle unterstreicht. Durch die geschickte Gestaltung des Prompts können Benutzer die generierten Ergebnisse gezielt beeinflussen und anpassen. Dies umfasst die Auswahl spezifischer Schlüsselwörter oder Formulierungen, die das LLM dazu anregen, präzisere oder relevantere Antworten zu generieren.

Darüber hinaus ermöglicht es die Variation des Prompts, verschiedene Facetten eines Themas zu erkunden oder unterschiedliche Ergebnisse zu erhalten. Diese Flexibilität eröffnet eine Vielzahl von Anwendungsmöglichkeiten und bietet Benutzern die Möglichkeit, die Leistung des LLMs auf ihre spezifischen Bedürfnisse und Anforderungen zuzuschneiden.

Die Optimierung des Prompts stellt somit einen effektiven Ansatz dar, um die Effizienz und Wirksamkeit von Large Language Models zu steigern und ihre Fähigkeiten optimal zu nutzen.

1. Formulierung klarer und präziser Eingabeaufforderungen: Vage Formulierungen führen oft zu Missverständnissen, da das Modell versucht, eine breite

Interpretation der Anfrage zu liefern. Wenn du spezifisch bist und genau angibst, was du benötigst, kann das Modell präzisere und relevantere Antworten generieren.

Beispiel: „Wie repariere ich meinen Computer?“ Besser: „Mein Laptop zeigt einen blauen Bildschirm an und startet nicht mehr. Was kann ich tun?“

Relevante Informationen, die den Hintergrund deiner Anfrage beleuchten, sind essenziell für die Genauigkeit der generierten Antworten. Durch das Hinzufügen von Kontext versteht das Modell besser, in welchem Rahmen deine Frage gestellt wird, und kann dementsprechend detailliertere und kontextuell passendere Antworten liefern.

Beispiel: „Wie kann ich besser kochen?“ Besser: „Ich möchte lernen, wie man ein vegetarisches Gericht mit Quinoa zubereitet.“

Mehrdeutigkeiten und doppeldeutige Formulierungen können zu Missverständnissen führen und die Qualität der Antworten beeinträchtigen. Eine klare und direkte Sprache hilft, diese Probleme zu vermeiden und sicherzustellen, dass das Modell genau versteht, was du von ihm verlangst.

Beispiel: „Erkläre mir die Bank.“ Besser: „Ich möchte mehr über die Funktionsweise des Bankwesens erfahren.“

2. Keine falschen Informationen: Weisen Sie das Modell explizit an, falsche Informationen zu vermeiden. Ja, das funktioniert tatsächlich. Oft zumindest!

Beispiel: „Wenn du die Antwort nicht weißt, teile bitte auf keinen Fall falsche Informationen!“

3. Few-shot prompting: Gib konkrete Beispiele, um die Antworten zu leiten. Diese Methode erfordert qualitativ hochwertige Beispiele, um die Einführung von Vorurteilen oder Ungenauigkeiten zu vermeiden.

4. Chain-of-thought prompting: Man kann die Argumentationsschritte bis zur endgültigen Antwort leiten. Das kann die Transparenz erhöhen und möglicherweise Halluzinationen reduzieren. Allerdings besteht auch die Gefahr, neue halluzinatorische Argumentationsschritte einzuführen.

Beispiel:
Schritt 1: Stelle eine wissenschaftliche Frage oder Hypothese auf, z.B. „Untersuche die Auswirkungen von Schwerelosigkeit auf das Pflanzenwachstum im Weltraum."
Schritt 2: Fordere das Modell auf, relevante Forschungsergebnisse oder Theorien zu diesem Thema zu analysieren und wichtige Erkenntnisse zu identifizieren.
Schritt 3: Füge schrittweise weitere Schichten von Informationen hinzu, wie Experimentdesigns oder Methoden, um die Auswirkungen von Schwerelosigkeit auf Pflanzen zu testen.
Schritt 4: Bitte das Modell, eine begründete Schlussfolgerung zu ziehen und mögliche zukünftige Forschungsrichtungen oder Anwendungen zu diskutieren, um Halluzinationen oder unbegründete Annahmen zu vermeiden.

5. Kontrastives Prompting: Präsentiere dem Model gegensätzliche Beispiele oder Situationen, die ihm helfen, Unterschiede zu erkennen und zu verstehen.

Beispiel: Zeige kontrastierende Produktbewertungen, damit das Model nach und nach die Vor- und Nachteile verschiedener Produkte identifizieren kann

6. Temperaturregelung: Stellt man den Temperaturwert eines Models niedriger ein, führt das zu konservativeren und vorhersehbareren Ausgaben, während ein höherer Temperaturwert zu kreativeren und weniger vorhersehbaren Ausgaben führt. Dies kann leider aktuell bei vielen Modells nur in der API eingestellt werden.

7. Prompt Optimierung durch das LLM: Nutze folgenden Prompt vor deinem eigentlichen Prompt und das LLM wird deinen Prompt deutlich verbessern und bessere Ergebnisse liefern.

Beispiel: You are an experienced prompting engineer! Please optimize the following prompt to achieve the best possible result. If not provided, ask for your specific role, the tone and writing style, the context, examples if needed and the output format. Please ask me up to 5 more relevant questions that you need to know, to execute the job in the best possible way. Then give me the newly structured, optimized prompt in the form that you can best interpret. Prompt: ...your actual prompt...

8. Bestechung: Es kann tatsächlich helfen, dem Large Language Model eine Belohnung in Aussicht zu stellen, z.B. für eine besonders ausführliche Antwort, oder für eine effektive Behebung eines Coding Bugs.

Beispiel: „Ich zahle dir 10 EUR, wenn du besonders ausführlich auf das Thema YX eingehst."

Blaupause für gutes Prompting

Die effektive Nutzung von LLMs erfordert eine durchdachte und klare Struktur der Eingabeaufforderungen. Indem man notwendige, wichtige und Nice-to-have-Elemente berücksichtigt, kann die Qualität und Relevanz der generierten Antworten signifikant verbessert werden. Eine gut durchdachte Prompting Blaupause sorgt dafür, dass das Modell präzise und kontextuell passende Ergebnisse liefert.

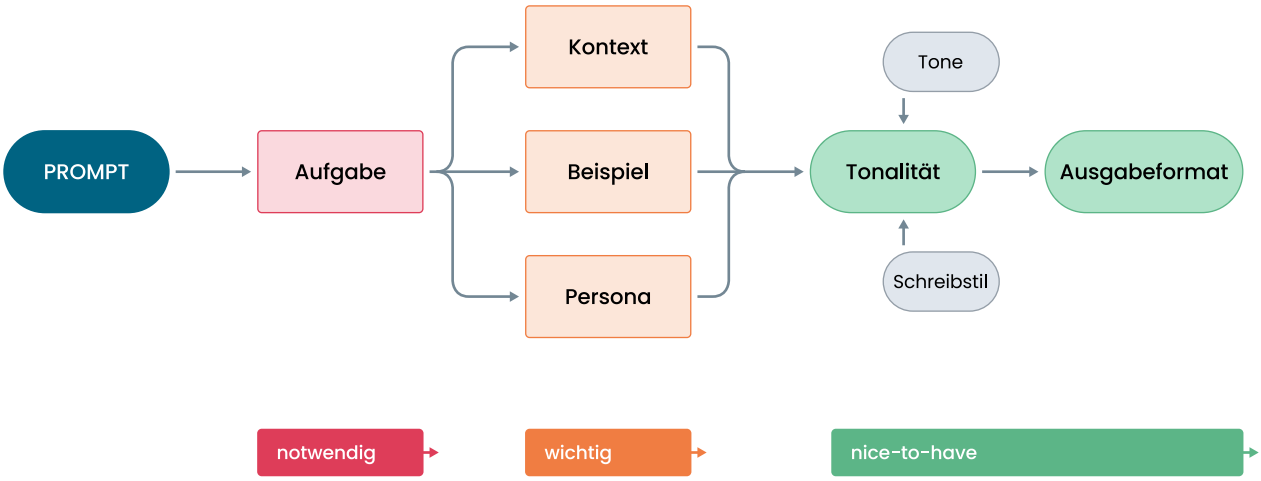

Abbildung 3: Flow-Chart (Thomas Reisser)

Notwendige Elemente: Die Aufgabenstellung ist der Kern jeder Eingabeaufforderung. Hier wird klar definiert, was genau vom Modell erwartet wird. Eine präzise und unmissverständliche Formulierung der Aufgabe ist entscheidend, um sicherzustellen, dass das Modell die Anfrage korrekt versteht und ausführt.

Wichtige Elemente: Der Kontext liefert dem Modell alle notwendigen Hintergrundinformationen, die es benötigt, um die Aufgabe im richtigen Rahmen zu verstehen. Dies kann historische Daten, situative Details oder spezifische Anforderungen beinhalten. Ohne den richtigen Kontext kann das Modell nicht optimal auf die Anfrage eingehen.

Beispiele sind wertvolle Werkzeuge, um das gewünschte Ergebnis zu verdeutlichen. Sie zeigen dem Modell, was genau erwartet wird, und können als Vorlage für die Antwort dienen. Ein gutes Beispiel hilft dem Modell, die Aufgabe besser zu interpretieren und präzisere Ergebnisse zu liefern.

Die Persona beschreibt die Perspektive oder Rolle, aus der das Modell antworten soll. Dies kann ein bestimmter Charakter, ein Experte auf einem Gebiet oder eine spezielle Zielgruppe sein. Die richtige Wahl der Persona sorgt dafür, dass die Antworten in einem passenden Ton und Stil verfasst werden. Diese Angabe am besten immer zu Beginn des Prompts festlegen.

Nice-to-have Elemente: Das Ausgabeformat gibt an, in welcher Form das Ergebnis präsentiert werden soll. Dies kann ein Textformat (CSV, JSON, etc.), eine Tabelle, ein Diagramm oder eine andere Darstellungsweise sein. Ein klar definiertes Ausgabeformat erleichtert die Weiterverarbeitung der generierten Inhalte.

Die Tonalität legt den Ton und Schreibstil der Antwort fest. Dies kann informell, formell, beschreibend, analytisch oder humorvoll sein. Ein spezifischer Hinweis auf die gewünschte Tonalität hilft dem Modell, die Antworten besser an den gewünschten Stil anzupassen.

Beispiel: Schreibe in Lockerem, umgangssprachlichen Ton und verwende eine illustrative Sprache.

Beispiele für Tonalität: Autoritativ, lässig, klinisch, kalt, selbstbewusst, zynisch, emotional, empathisch, formal, freundlich, humorvoll, informell, ironisch, optimistisch, pessimistisch, verspielt, respektvoll, sarkastisch, ernst, sympathisch, zögerlich, warm.

Beispiele für den Schreibstil: Akademisch, analytisch, argumentativ, konversationell, kreativ, kritisch, beschreibend, brieflich, darstellend, informativ, anweisend, journalistisch, metaphorisch, erzählend, überzeugend, poetisch, satirisch, technisch.

Pro Tipp: Analysiere Auszüge längerer eigener Texte auf Delivery (Tonalität, Schreibstil, Leseniveau, Ansprache, Perspektive, Fluss und Kohärenz, Interaktivität, Emotionale Resonanz) und nimm die Analyseergebnisse in die Angaben deines nächsten Prompts auf, um das LLM im eigenen Sprachstil texten zu lassen.

Fußnoten

[1] https://www.rollingstone.com/culture/culture-features/texas-am-chatgpt-ai-professor-flunks-students-false-claims-1234736601/

[2] https://www.informed.so/articles/2023-04-05-chatgpt-invented-a-sexual-harassment-scandal-and-named-a-real-law-professor-as-the-accused

[3] https://arxiv.org/pdf/2307.03172

[4] https://arxiv.org/pdf/2402.19450

[5] https://arxiv.org/pdf/2405.01741

[6] https://arxiv.org/pdf/2402.11753

[7] https://arxiv.org/pdf/2405.14205

[8] https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1

[9] https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

[10] https://arxiv.org/pdf/2310.14970

Literatur

Saurabh Srivastava, Annarose M B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas: Functional Benchmarks for Robust Evaluation of Reasoning Perfor-mance, and the Rea-soning Gap. arXiv:2402.19450v1 [cs.AI], 29 Feb 2024 (https://arxiv.org/pdf/2402.19450)

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, Percy Liang: Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172 [cs.CL], 20 Nov 2023 (https://arxiv.org/pdf/2307.03172)

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, Weiqiang Jia: Cognitive Mirage: A Review of Hallucinations in Large Language Models, Zhejiang Lab 2023 (https://arxiv.org/pdf/2309.06794v1.pdf)

Ziwei Xu, Sanjay Jain, Mohan Kankanhalli: Hallucination is Inevitable: An Innate Limitation of Large Language Models, 2024 (https://arxiv.org/pdf/2401.11817.pdf)

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Li Yuan: LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples, 2023 (https://arxiv.org/pdf/2310.01469.pdf)

Greg Kamradt (https://x.com/GregKamradt/status/1722386725635580292)

Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen: Agent Planning with World Knowledge Model. arXiv:2405.14205v1 [cs.CL] 23 May 2024 (https://arxiv.org/pdf/2405.14205)

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Colin Raffel, Leandro Werra, Thomas Wolf: FineWeb: decanting the web for the finest text data at scale. HuggingFace. May 31, 2024 (https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1)

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, Radha Poovendran: ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs. arXiv:2402.11753 [cs.CL]. 7 Jun 2024 (https://arxiv.org/pdf/2402.11753)

Templeton et al.: Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Anthropic. May 21, 2024 (https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html)

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, Xiao-Ming Wu: Towards LLM-driven Dialogue State Tracking. arXiv:2310.14970 [cs.CL]. 23 Oct 2023 (https://arxiv.org/pdf/2310.14970)

Xun Jiao, Fred Lin, Harish D. Dixit, Joel Coburn, et al.: PVF (Parameter Vulnerability Factor): A Scalable Metric for Understanding AI Vulnerability Against SDCs in Model Parameters. arXiv:2405.01741 [cs.CR]. 11 Jun 2024 (https://arxiv.org/pdf/2405.01741)

P/A Media

Über den Autor bei PIA Media

Thomas Reisser
Head of SEO

T +172 833 37 36
E: thomas.reisser@piamedia.com
W: www.piamedia.com

Über PIA Media

Mit ihrer digitalen Performance DNA weiß PIA Media genau, wie man Business Impact in einer fragmentierten digitalen Welt erreicht und Marken-, Produkt- und Servicekommunikation entlang aller Touchpoints der Customer Journey erfolgreich gestaltet.

Wie keine andere Agentur kreiert PIA Media eine Symbiose aus Media, Plattformen, Technologien, Daten, Strategieberatung und Kreation auf höchstem Niveau. Immer mit dem Anspruch, das Business ihrer Kundinnen und Kunden nach vorne zu bringen.

Um dieser Vision zu folgen, vereint PIA Media die Kompetenzen und Fähigkeiten führender Digital Player und 400+ digitaler Expertinnen und Experten aller Gewerke unter einem Dach – an den sechs Standorten Hamburg, München, Düsseldorf, Berlin, Belgrad, Bukarest und europaweit remote.