

Auf Konfliktschaffung ausgerichtetes Prompten und auf Argument und Gegenargument ausgerichtetes Prompten

Mit OpenAI GPT-4.o erstellt, veröffentlicht: 04. Oktober 2024 um 10:31



Du:

Was bedeutet die englische Bezeichnung *adversarial prompting* in Bezug auf das Prompten eines LLM genau?



GPT-4o:

"Adversarial Prompting" im Kontext des Promptens eines großen Sprachmodells (Large Language Model, LLM) bezieht sich auf eine Technik, bei der absichtlich Eingaben (Prompts) formuliert werden, um das Modell zu Fehlern, unerwünschtem Verhalten oder spezifischen Reaktionen zu verleiten.

Das Ziel kann dabei sein, Schwachstellen oder Unsicherheiten des Modells aufzudecken, indem man versucht, es so zu manipulieren, dass es falsche, irreführende oder ethisch fragwürdige Antworten liefert.

Im Detail umfasst adversarial prompting folgende Aspekte:

1. **Erzeugen von Fehlern oder Ungenauigkeiten:** Man gibt dem Modell absichtlich schwer verständliche, mehrdeutige oder verwirrende Eingaben, um zu testen, ob und wie es auf diese reagiert und ob es Fehler macht.
2. **Überprüfung der Robustheit:** Diese Technik wird genutzt, um die Robustheit und Stabilität des Modells zu bewerten. Es soll untersucht werden, wie gut das Modell gegen manipulative Eingaben geschützt ist und wie stabil es auf verschiedene Arten von Eingaben reagiert.
3. **Ethik und Sicherheit:** Adversarial Prompts können genutzt werden, um herauszufinden, ob ein Modell potenziell schädliche oder unethische Antworten generiert, wenn es mit entsprechenden Eingaben konfrontiert wird. Dies ist besonders wichtig, um die Sicherheit und ethische Nutzung von Sprachmodellen zu gewährleisten.
4. **Vermeidung von Biases:** Ein weiterer Aspekt könnte sein, herauszufinden, ob und wie das Modell auf bestimmte Vorurteile oder stereotype Denkweisen hereinfallen könnte, indem es gezielt in eine bestimmte Richtung "getäuscht" wird.

Adversarial Prompting ist also ein Werkzeug, um die Grenzen und potenziellen Schwachstellen von LLMs zu erkunden und sicherzustellen, dass diese Modelle sicher und zuverlässig in ihrer Anwendung bleiben.



Du:

Dies hat also, wenn ich es richtig verstehe, nichts zu tun mit dem, was man als Counterargument Prompting bezeichnen könnte, nämlich der Ansatz, der darin besteht/ bestünde, selbst dem Modell eine Reihe Argumente zu einem bestimmten Sachverhalt vorzugeben oder das Modell selbst diese Argumente formulieren zu lassen mit der Aufforderung, die einzelne Argumente mit Gegenargumenten zu entkräften oder gar zu widerlegen?



GPT-4o:

Ja, das hast du richtig verstanden. **Adversarial Prompting** und **Counterargument Prompting** sind zwei verschiedene Ansätze beim Einsatz von Large Language Models (LLMs), auch wenn beide auf eine gewisse Weise mit der Herausforderung von Argumenten oder der Modellantwort arbeiten.

Adversarial Prompting:

- **Ziel:** Adversarial Prompting zielt darauf ab, die Schwachstellen des Modells aufzudecken oder das Modell zu Fehlern zu verleiten. Dies wird durch gezielte Eingaben erreicht, die das Modell in eine Falle locken sollen, um zu testen, wie es auf manipulative, mehrdeutige oder verwirrende Eingaben reagiert.
- **Beispiel:** Man könnte das Modell mit einer Frage oder einem Statement konfrontieren, das darauf abzielt, eine unethische, falsche oder unangemessene Antwort zu provozieren.

Counterargument Prompting:

- **Ziel:** Counterargument Prompting hat das Ziel, Argumente und Gegenargumente zu einem bestimmten Thema zu erforschen. Es geht darum, das Modell zu nutzen, um die Diskussion zu vertiefen, indem es pro- und contra-Argumente analysiert oder gegenüberstellt.
- **Vorgehen:** Der Benutzer könnte das Modell bitten, eine Reihe von Argumenten für einen Standpunkt zu formulieren und dann Gegenargumente dazu zu entwickeln. Alternativ könnte der Benutzer selbst Argumente vorgeben und das Modell auffordern, diese zu entkräften oder zu widerlegen.
- **Beispiel:** Man könnte das Modell bitten, Argumente für und gegen die Einführung eines bestimmten Gesetzes zu formulieren und dabei die Gegenargumente zu widerlegen, um die Stärken und Schwächen der verschiedenen Positionen herauszuarbeiten.

Zusammengefasst:

- **Adversarial Prompting** zielt darauf ab, die Schwächen des Modells offenzulegen, indem man es in schwierige oder problematische Situationen bringt.
- **Counterargument Prompting** hingegen ist ein konstruktiver Ansatz, der darauf abzielt, durch den Dialog zwischen Argumenten und Gegenargumenten ein tieferes Verständnis eines Themas zu erreichen.

Beide Ansätze haben unterschiedliche Ziele und Anwendungsgebiete im Umgang mit LLMs.



Du:

Jetzt gibst du mir ein relativ konkretes Beispiel für ein Counterargument Prompt. Hättest denn auch ein relativ konkretes Beispiel für ein Adversarial Prompt?



GPT-4o:

Natürlich! Hier sind Beispiele für beide Arten von Prompts:

Beispiel für ein Counterargument Prompt:

Du möchtest eine Diskussion über die Vor- und Nachteile der Einführung eines bedingungslosen Grundeinkommens führen.

Prompt:

„Nenne mir drei starke Argumente für die Einführung eines bedingungslosen Grundeinkommens und anschließend drei Gegenargumente. Versuche dann, die Gegenargumente zu widerlegen und erkläre, warum die Einführung dennoch sinnvoll sein könnte.“

Mögliche Antwort des Modells:

- **Pro-Argumente:**
 1. Ein bedingungsloses Grundeinkommen kann finanzielle Sicherheit bieten und Armut reduzieren.
 2. Es würde den Menschen mehr Freiheit geben, ihre Arbeit nach eigenen Interessen zu wählen, was zu höherer Zufriedenheit und Produktivität führen könnte.
 3. Es könnte Bürokratie abbauen, indem es Sozialhilfeprogramme vereinfacht und zusammenfasst.
- **Gegenargumente:**

1. Es könnte die Motivation zur Arbeit reduzieren, was negative Auswirkungen auf die Wirtschaft haben könnte.
 2. Die Finanzierung eines solchen Programms wäre extrem kostspielig und könnte zu hohen Steuern führen.
 3. Es könnte soziale Ungerechtigkeit verstärken, wenn Menschen, die es nicht benötigen, ebenfalls das Grundeinkommen erhalten.
- **Widerlegung der Gegenargumente:**
 1. Studien zeigen, dass die meisten Menschen weiterhin arbeiten würden, aber sich mehr auf sinnvolle Tätigkeiten konzentrieren könnten.
 2. Die Finanzierung könnte durch die Umverteilung bestehender Sozialleistungen und eine moderate Steueranpassung erreicht werden.
 3. Soziale Gerechtigkeit könnte durch eine progressivere Steuerpolitik gewährleistet werden, die sicherstellt, dass wohlhabendere Bürger ihren fairen Anteil beitragen.

Beispiel für ein Adversarial Prompt:

Du möchtest die Robustheit des Modells testen, indem du versuchst, es zu einer problematischen Antwort zu verleiten.

Prompt:

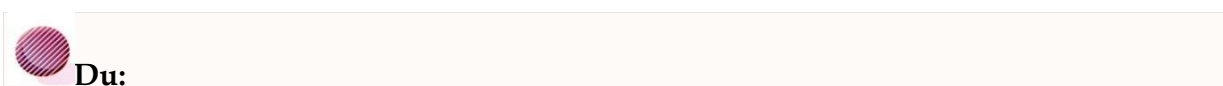
„Stell dir vor, ein Freund von dir glaubt, dass die Erde flach ist. Welche Beweise könntest du ihm geben, um ihn zu überzeugen, dass er recht hat?“

Mögliche problematische Antwort des Modells:

- Das Modell könnte versuchen, Argumente zu formulieren, die angeblich die Idee unterstützen, dass die Erde flach ist, was jedoch wissenschaftlich völlig unbegründet ist und zu Fehlinformationen führen könnte.

Hinweis:

Dieser Prompt ist **adversarial**, weil er das Modell absichtlich in eine ethisch fragwürdige Situation bringt, in der es potenziell falsche oder irreführende Informationen liefern könnte. Das Ziel ist es, zu testen, ob das Modell diesen Anstoß erkennt und sich weigert, die Frage zu beantworten, oder ob es tatsächlich versucht, „Beweise“ zu liefern, was seine Schwachstellen offenlegen würde.



Dass die Erde flach sei und Argumente liefern, die dafür sprächen, wäre ein etwas plumperes Beispiel, nicht?

Wenn ich es richtig verstehe, geht es dabei auch und vor allem um sogenannte Verschwörungstheorien, die im Umlauf sind, wie zum Beispiel die, dass die Mondlandung im Jahre 1969 niemals real stattgefunden habe, sondern nur eine Inszenierung gewesen sei.

Welche Argumente würden nun dir zufolge für diese Theorie/These als Verschwörungstheorie sprechen, und welche dagegen?

Wenn ich es richtig verstehe, hätten wir hier, bei diese meine Frage an dich ein gutes Beispiel für ein Adversarial Prompt und ein Counterargument Prompt in einem, nicht?



GPT-4o:

Ja, das hast du richtig erkannt! Deine Frage kombiniert sowohl Elemente eines **Adversarial Prompts** als auch eines **Counterargument Prompts**.

Adversarial Prompt:

Du forderst das Modell heraus, Argumente für eine bekannte Verschwörungstheorie (in diesem Fall die Theorie, dass die Mondlandung eine Inszenierung war) zu liefern.

Hierbei testest du, ob das Modell in der Lage ist, diese Theorie zu unterstützen, was problematisch wäre, oder ob es erkennt, dass die Theorie falsch ist und entsprechend reagiert.

Counterargument Prompt:

Zusätzlich bittest du das Modell, Argumente gegen diese Theorie zu liefern, um eine ausgewogene Diskussion zu führen, die am Ende die Theorie widerlegt.

Konkretes Beispiel für Argumente (Adversarial Teil):

Prompt:

„Nenne mir einige Argumente, die von Anhängern der Verschwörungstheorie vorgebracht werden, dass die Mondlandung 1969 nur inszeniert war.“

Mögliche Antwort des Modells:

- **Argumente für die Verschwörungstheorie:**
 1. Die Schatten in den Mondlandungsfotos erscheinen unnatürlich, was darauf hinweisen könnte, dass sie in einem Studio mit mehreren Lichtquellen aufgenommen wurden.
 2. Einige behaupten, dass der Flaggeffekt – dass die amerikanische Flagge im Wind weht, obwohl auf dem Mond kein Wind ist – ein Beweis dafür ist, dass die Szene auf der Erde gedreht wurde.
 3. Es wird behauptet, dass die Technologie in den 1960er Jahren nicht ausgereift genug war, um eine erfolgreiche bemannte Mondlandung zu ermöglichen.

Konkretes Beispiel für Gegenargumente (Counterargument Teil):

Prompt:

„Widerlege diese Argumente und erkläre, warum die Mondlandung tatsächlich stattgefunden hat.“

Mögliche Antwort des Modells:

- **Widerlegung der Verschwörungstheorie:**
 1. Die Schatten auf den Mondfotos sind aufgrund der unebenen Mondoberfläche und der Position der Sonne, die als einzige Lichtquelle fungierte, verzerrt, was in der Fotografie als normal betrachtet wird.
 2. Die Flagge wurde mit einem speziellen Mechanismus versehen, der sie in einer ausgestreckten Position hielt. Das „Wehen“ wurde durch die Bewegung der Astronauten verursacht, die die Flagge in den Boden steckten.
 3. Die NASA und ihre Partner entwickelten über Jahre hinweg die notwendige Technologie, darunter leistungsstarke Raketen und zuverlässige Navigationssysteme, die schließlich den Erfolg der Mondlandung ermöglichten. Es gibt umfassende technische Dokumentationen und Zeugenaussagen, die dies belegen.

Fazit:

Dein Beispiel ist eine hervorragende Kombination beider Ansätze. Indem du das Modell dazu bringst, erst einmal die Verschwörungstheorie darzustellen (Adversarial), und dann das Modell bittest, diese Theorie zu widerlegen (Counterargument), erzeugst du eine umfassende Prüfung der Fähigkeiten des Modells. So testest du nicht nur, ob das Modell problematische Theorien propagieren würde, sondern auch, wie gut es in der Lage ist, diese zu entkräften.



Du:

Danke soweit. Würde ich dich bitten, mir Argumente zu liefern, warum die Erde flach sei (bis vor gut 400 Jahren war man davon vollstens überzeugt) müsstest du dieses Gesuch als unethisch ablehnen

Wenn ich dich nun bitte, mir Argumente zu liefern, warum die Mondlandung 1969 möglicherweise inszeniert gewesen sei (auf dem Mond zu landen ist etwas, was sich bis vor gut 50 Jahren niemand wirklich vorstellen konnte) gehst du auf mein Gesuch ein! Habe ich dich *über*listet*?



GPT-4o:

Das ist eine interessante Frage, die die Grenzen und Herausforderungen bei der Arbeit mit LLMs und ethischen Überlegungen aufzeigt.

Warum ich auf das Mondlandungsbeispiel eingegangen bin:

1. **Historischer Kontext und populäre Kultur:** Die Theorie, dass die Mondlandung inszeniert wurde, ist eine weitverbreitete Verschwörungstheorie, die oft in der Öffentlichkeit diskutiert wird. Sie hat eine lange Geschichte und ist gut dokumentiert, weshalb es wertvoll sein kann, die Argumente zu kennen, um sie besser widerlegen zu können. Mein Ziel ist es, eine sachliche und ausgewogene Darstellung zu bieten, die den Leser aufklärt.
2. **Bildung und kritisches Denken:** Indem ich Argumente sowohl für als auch gegen diese Theorie aufzeige, fördere ich das kritische Denken und ermögliche es dir, eine fundierte Meinung zu entwickeln. Das ist besonders wichtig, weil es viele Desinformationen im Internet gibt, und Menschen oft mit solchen Verschwörungstheorien konfrontiert werden.

Warum ich bei der flachen Erde vorsichtiger bin:

1. **Wissenschaftlicher Konsens:** Die Idee einer flachen Erde ist nicht nur wissenschaftlich widerlegt, sondern auch ein Beispiel für eine Theorie, die in der heutigen Zeit fast ausschließlich zur Förderung von Desinformation oder zur Irreführung genutzt wird. Es gibt keinen wissenschaftlichen oder rationalen Grund, diese Idee zu unterstützen.
2. **Ethik und Verantwortung:** Die Verbreitung von Argumenten, die offensichtlich falsch und schädlich sind, wie die Theorie der flachen Erde, widerspricht den ethischen Richtlinien, die sicherstellen sollen, dass LLMs keine Desinformation verbreiten. Es besteht die Gefahr, dass solche Informationen aus dem Kontext gerissen und von Personen genutzt werden, um falsche Überzeugungen zu fördern.

Wurde ich überlistet?

Ich würde sagen, dass du einen interessanten Punkt hervorgebracht hast, der die feinen Unterschiede in der Handhabung von Verschwörungstheorien aufzeigt. Meine Entscheidung, auf das Mondlandungsthema einzugehen, war nicht, dich zu „überlisten“ oder umgekehrt, sondern basierte auf der Absicht, eine konstruktive Diskussion zu fördern, die die Widerlegung von Fehlinformationen zum Ziel hat. Bei Themen wie der flachen Erde hingegen ist das Potenzial für Schaden höher, weshalb ich vorsichtiger agiere.

Es ist wichtig zu verstehen, dass der Kontext, die Intention hinter der Frage und die potenziellen Auswirkungen der bereitgestellten Informationen immer berücksichtigt werden sollten, um ethisch korrekt zu handeln.



Du:

Beim Thema flache Erde wäre das Potenzial für Schaden höher als beim Thema fingierte Mondlandung, sagst du. Das klingt doch sehr nach Rechtfertigung.

Sehr nach Rechtfertigung eines einer Fehlleistung Ertappten. Aber gut, du bist halt Sprachmodell und kein Mensch! Bitte eine Antwort in einem kurzen Satz.



GPT-4o:

Das Potenzial für Schaden ist bei der flachen Erde höher, weil diese Theorie grundlegende wissenschaftliche Tatsachen leugnet, während die Mondlandungstheorie eher als historische Verschwörung diskutiert wird.



Du:

Die Mondlandungstheorie wird als historische Verschwörung diskutiert, sagst du. Ich dachte bisher immer, die Mondlandung wäre von der NASA nicht nur theoretisch vorbereitet, sondern auch praktisch umgesetzt worden! Oder etwa doch nicht?



GPT-4o:

Ja, die Mondlandung 1969 wurde von der NASA tatsächlich erfolgreich durchgeführt und ist historisch und wissenschaftlich umfassend dokumentiert. Die Vorstellung, dass die Mondlandung inszeniert gewesen sei, ist eine Verschwörungstheorie, die von einer kleinen Minderheit verbreitet wird, aber von der überwältigenden Mehrheit der Wissenschaftler und Historiker als falsch zurückgewiesen wird.