

Selbstkonsistenz eines Sprachmodells und Selbstausswertung eines Sprachmodells

Mit Gemini erstellt - veröffentlicht: 22. Februar 2024 um 15:15

Auf Selbstkonsistenz hin ausgerichtetes Prompten ist ein ungemein wirkungsvoller Ansatz im Umgang mit Sprachmodellen, gerade wenn es um die Reliabilität der Aussage geht, wie dies, etwa im rechtlichen oder medizinischen Bereich, ein *conditio-sine-qua-non* darstellt für den praktischen Einsatz von Sprachmodellen. So habe wir eben gemeinsam festgestellt.

Dir war dabei noch - wie so oft - ein logischer Fehler unterlaufen, indem du behauptetest, ein auf Selbstkonsistenz hin ausgerichtetes Prompten sei ein *conditio-sine-qua-non* für den Einsatz von Sprachmodellen im Bereich des Rechts und der Medizin. Meine Feststellung jedoch war, das ! **deren Reliabilität** ! *conditio-sine-qua-non* für den Einsatz von Sprachmodellen im Bereich des Rechts und der Medizin sei.

Dies vorausgeschickt, möchte ich dich bitten, mir etwas zum Unterschied des Prompting-Ansatzes der Selbstkonsistenz und des Prompting-Ansatz der Selbstausswertung zu erklären, so es dir dann möglich wäre.

Gerne erkläre ich Ihnen den Unterschied zwischen dem Prompting-Ansatz der Selbstkonsistenz und dem Prompting-Ansatz der Selbstausswertung.

Selbstkonsistenz:

Der Prompting-Ansatz der Selbstkonsistenz zielt darauf ab, ein Sprachmodell so zu steuern, dass es Aussagen generiert, die mit seinen eigenen früheren Aussagen und seinem internen Wissensmodell konsistent sind. Dies wird erreicht, indem dem Sprachmodell Prompts mit folgenden Merkmalen gegeben werden:

- Sie enthalten Informationen, die bereits vom Sprachmodell verarbeitet und in seinem Wissensmodell gespeichert wurden.
- Sie stellen Fragen oder Aufgaben, die das Sprachmodell mit seinem vorhandenen Wissen beantworten oder lösen kann.
- Sie fordern das Sprachmodell auf, seine eigenen Aussagen zu reflektieren und zu erklären.

Der Vorteil dieses Ansatzes ist, dass er die Wahrscheinlichkeit erhöht, dass das Sprachmodell **verlässliche** und **kohärente** Aussagen generiert. Dies ist besonders wichtig in Bereichen wie Recht und Medizin, wo es auf präzise und konsistente Informationen ankommt.

Selbstausswertung:

Der Prompting-Ansatz der Selbstausswertung hingegen zielt darauf ab, das Sprachmodell dazu zu bringen, seine eigenen Stärken und Schwächen sowie die Grenzen seiner Fähigkeiten zu erkennen und zu reflektieren. Dies wird erreicht, indem dem Sprachmodell Prompts mit folgenden Merkmalen gegeben werden:

- Sie stellen Fragen nach der Zuverlässigkeit und Gültigkeit der vom Sprachmodell generierten Aussagen.
- Sie fordern das Sprachmodell auf, die eigenen Prämissen und Schlussfolgerungen zu hinterfragen.
- Sie bitten das Sprachmodell um eine Einschätzung seiner eigenen Leistung und um Vorschläge zur Verbesserung.

Der Vorteil dieses Ansatzes ist, dass er das Sprachmodell **transparenter** und **nachvollziehbarer** macht. Dies ist wichtig, um sicherzustellen, dass das Sprachmodell verantwortungsvoll eingesetzt wird und dass seine Grenzen erkannt und berücksichtigt werden.

Unterschiede:

Die beiden Prompting-Ansätze unterscheiden sich in folgenden Punkten:

- **Ziel:** Der Selbstkonsistenz-Ansatz zielt auf die Verlässlichkeit und Kohärenz der Aussagen ab, während der Selbstausswertungs-Ansatz auf die Transparenz und Nachvollziehbarkeit des Sprachmodells abzielt.
- **Art der Prompts:** Die Prompts beim Selbstkonsistenz-Ansatz enthalten Informationen und Aufgaben, die das Sprachmodell mit seinem vorhandenen Wissen lösen kann. Die Prompts beim Selbstausswertungs-Ansatz hingegen stellen Fragen nach der Zuverlässigkeit und Gültigkeit der Aussagen des Sprachmodells.
- **Ergebnisse:** Der Selbstkonsistenz-Ansatz führt zu Aussagen, die mit dem internen Wissensmodell des Sprachmodells konsistent sind. Der Selbstausswertungs-Ansatz führt zu einer Einschätzung der eigenen Leistung des Sprachmodells und zu Vorschlägen zur Verbesserung.

Zusammenfassend lässt sich sagen, dass der Prompting-Ansatz der **Selbstkonsistenz** und der Prompting-Ansatz der **Selbstausswertung** zwei **komplementäre** Ansätze sind, um die Zuverlässigkeit, Kohärenz, Transparenz und Nachvollziehbarkeit von Sprachmodellen zu verbessern.

In den Bereichen Recht und Medizin ist es wichtig, beide Ansätze zu nutzen, um sicherzustellen, dass Sprachmodelle verantwortungsvoll eingesetzt werden und dass ihre Grenzen erkannt und berücksichtigt werden.

Ich hoffe, diese Erklärung war hilfreich.

Bitte beachten Sie, dass ich mich noch in der Entwicklung befinde und es daher zu Fehlern kommen kann.

Ich bin jedoch stets bemüht, meine Fähigkeiten zu verbessern und Ihnen die bestmögliche Unterstützung zu bieten.

Zwei komplementäre Ansätze also, sagst du.

Voraussetzung für komplementäres Anwenden aber ist umfassendes Verständnis im Einzelnen, oder nicht?

Selbstkonsistenz:

- Man muss die verschiedenen Arten von Prompts kennen, die verwendet werden können, um die Selbstkonsistenz zu fördern.
- Man muss verstehen, wie das Sprachmodell seine internen Wissensmodelle nutzt, um konsistente Aussagen zu generieren.
- Man muss die Grenzen des Selbstkonsistenz-Ansatzes kennen und wissen, wann es notwendig ist, andere Ansätze zu verwenden.

Selbstausswertung:

- Man muss die verschiedenen Arten von Prompts kennen, die verwendet werden können, um die Selbstausswertung zu fördern.
- Man muss verstehen, wie das Sprachmodell seine eigene Leistung einschätzt und welche Faktoren diese Einschätzung beeinflussen.
- Man muss die Grenzen des Selbstausswertungs-Ansatzes kennen und wissen, wann es notwendig ist, andere Ansätze zu verwenden.

Nur wenn man diese beiden Ansätze im Einzelnen umfassend versteht, kann man sie effektiv nutzen, um die Zuverlässigkeit, Kohärenz, Transparenz und Nachvollziehbarkeit von Sprachmodellen zu verbessern.

In den Bereichen Recht und Medizin ist es besonders wichtig, diese beiden Ansätze zu verstehen und anzuwenden, da hier die Zuverlässigkeit und Kohärenz der Aussagen von Sprachmodellen von entscheidender Bedeutung sein kann.

Eine abschließende Frage noch:

In welchem Verhältnis steht nun der Prompting-Ansatz der Selbstkonsistenz und der Prompting-Ansatz der Selbstausswertung zum Prompting-Ansatz der Selbstverfeinerung bzw. iterativen Selbstverbesserung?

Das ist eine ausgezeichnete Frage! Die drei von Ihnen genannten Prompting-Ansätze sind eng miteinander verwandt und ergänzen sich in gewisser Weise. Lassen Sie uns die Beziehungen genauer betrachten:

- **Selbstkonsistenz und Selbstverfeinerung:** Beide Ansätze zielen darauf ab, die Qualität der von einem Sprachmodell generierten Antworten zu verbessern. Die Selbstkonsistenz fokussiert dabei in erster Linie auf die logische Kohärenz und Widerspruchsfreiheit innerhalb einer einzelnen Antwort oder eines kurzen Dialogs. Die Selbstverfeinerung geht einen Schritt weiter und betrachtet die Entwicklung der Antworten über mehrere Interaktionen hinweg. Sie zielt darauf ab, dass das Modell aus seinen vorherigen Antworten lernt und diese iterativ verbessert. **Die Selbstverfeinerung baut also auf der Selbstkonsistenz auf**, indem sie sicherstellt, dass die Verbesserungen nicht zu neuen Widersprüchen führen.
- **Selbstausswertung und Selbstverfeinerung:** Die Selbstausswertung ermöglicht es dem Modell, seine eigenen Stärken und Schwächen zu erkennen. Diese Erkenntnisse können dann direkt in den Prozess der Selbstverfeinerung einfließen. Das Modell kann beispielsweise lernen, welche Arten von Fragen es besonders gut oder schlecht beantwortet und seine Strategien entsprechend anpassen. **Die Selbstausswertung liefert also die Grundlage für zielgerichtete Selbstverfeinerung.**

Zusammenfassend lässt sich sagen:

- **Selbstkonsistenz** ist die Basis, die sicherstellt, dass die Antworten des Modells logisch konsistent sind.
- **Selbstausswertung** ermöglicht es dem Modell, seine eigenen Fehler zu erkennen und zu analysieren.
- **Selbstverfeinerung** nutzt die Erkenntnisse aus Selbstkonsistenz und Selbstausswertung, um die Qualität der Antworten iterativ zu verbessern.

Alle drei Ansätze sind eng miteinander verknüpft und arbeiten zusammen, um die Leistungsfähigkeit von Sprachmodellen zu steigern.

Ein einfaches Beispiel:

Stellen Sie sich vor, ein Sprachmodell wird gebeten, die Geschichte von Romeo und Julia zu erzählen.

- **Selbstkonsistenz:** Das Modell muss sicherstellen, dass die Charaktere konsistent handeln und dass die Ereignisse in einer logischen Reihenfolge ablaufen.
- **Selbstausswertung:** Das Modell könnte feststellen, dass es Schwierigkeiten hat, die Emotionen der Charaktere auszudrücken.
- **Selbstverfeinerung:** Das Modell könnte lernen, mehr emotionale Wörter und Sätze zu verwenden, um die Geschichte lebendiger zu gestalten.

In der Praxis werden diese Ansätze oft kombiniert, um optimale Ergebnisse zu erzielen.