DATALYXT

GRUNDLAGEN

Warum Sprachmodelle Dinge erfinden

Halluzinationen lassen sich senken, aber nicht abschalten. Der Grund liegt im Verfahren selbst.

Eine Halluzination ist kein Programmfehler.

Das Modell arbeitet dabei genau so, wie es gebaut ist.

Was gemeint ist

Eine Ausgabe, die gut formuliert, richtig aufgebaut und sachlich falsch ist. Sie sieht aus wie jede andere Antwort, weil sie auf demselben Weg entstanden ist.

Beispiel: Gefragt nach dem Zahlungsziel in einem Vertrag antwortet das Modell „30 Tage“. Im Vertrag stehen 14. Nichts an der Antwort sieht falsch aus.

Fünf Ursachen, die zusammenwirken

1. **Das Modell schlägt nichts nach, es setzt fort.**
   Es erzeugt die wahrscheinlichste Fortsetzung eines Textes. Ob eine Aussage zutrifft, kommt darin nicht vor.

2. **Es gibt keine Wissensdatenbank im Modell.**
   Was es „weiß“, liegt verteilt in Milliarden Gewichten. Einzelne Angaben lassen sich dort nicht nachschlagen.

3. **Das Modell kennt seinen eigenen Kenntnisstand nicht.**
   Es unterscheidet nicht zwischen „das weiß ich“ und „das ergänze ich plausibel“.

4. **Raten wird belohnt, Zugeben nicht.**
   Wer nur nach richtig und falsch bewertet, belohnt jeden Versuch. Eine eingestandene Lücke bringt keine Punkte.

5. **Sprache ist mehrdeutig.**
   Eine Frage lässt oft mehrere Lesarten zu. Das Modell wählt eine, ohne die Wahl sichtbar zu machen.

Jede Modellgeneration halluziniert seltener als die vorige. Keine bringt die Rate auf null, weil die Ursache im Verfahren liegt und nicht in der Güte des einzelnen Modells.

DATALYXT

MASSNAHMEN

Sechs Maßnahmen, die wirken

Keine verhindert Halluzinationen. Alle sorgen dafür, dass ein falscher Wert auffällt, bevor er wirkt.

1 Nur aus mitgelieferten Quellen antworten

Die Suche liefert die Textstellen, das Modell formuliert daraus und nennt die Fundstelle.

Wirkung: Jede Aussage ist am Original nachprüfbar.

Grenze: Eine Fundstelle beweist nicht, dass richtig zusammengefasst wurde.

2 Ausgabeform festlegen und prüfen

Feste Felder, Datentypen und zulässige Werte. Was nicht passt, wird abgewiesen statt weitergereicht.

Wirkung: Erfundene Kategorien und Formfehler fallen sofort auf.

Grenze: Ein Wert kann formal richtig und inhaltlich falsch sein.

3 Nicht rechnen lassen

Zahlen, Fristen und Einstufungen berechnet das Programm. Das Modell liest nur die Ausgangswerte.

Wirkung: Die häufigste Fehlerklasse entfällt vollständig.

Grenze: Ein falsch ausgelesener Ausgangswert wirkt weiter.

4 Gegen eine unabhängige Quelle prüfen

Regelwerk, Stammdaten oder eine zweite Berechnung entscheiden, ob das Ergebnis Bestand hat.

Wirkung: Fehler werden erkannt, bevor sie weiterlaufen.

Grenze: Das Modell darf sich nicht selbst prüfen. Eigene Ausgaben bewertet es zu gut.

5 Eine feste Prüfmenge aufbauen

Reale Vorgänge mit bekannten Ergebnissen. Gemessen wird Feld für Feld, nicht als Gesamteindruck.

Wirkung: Änderungen an Prompt oder Modell werden messbar statt spürbar.

Grenze: Sie deckt nur ab, was in ihr vorkommt. Sie muss mitwachsen.

6 Einen Menschen an einer benannten Stelle

Festgelegt wird, welche Rolle an welchem Punkt entscheidet und wie das dokumentiert wird.

Wirkung: Die Verantwortung für das Ergebnis ist zugeordnet.

Grenze: Prüfen muss schneller gehen als Selbermachen, sonst wird es zur Formsache.

DATALYXT

GEGENSTÜCK

Was nicht hilft

Fünf Maßnahmen, die häufig vorgeschlagen werden und das Problem nicht lösen.

Die Anweisung „erfinde nichts“

Sie senkt die Rate ein wenig. Das Modell erkennt aber nicht, wann es gerade erfindet.

Das Modell nach seiner Sicherheit fragen

Die Selbsteinschätzung ist genau dort am wenigsten verlässlich, wo die Antwort falsch ist.

Ein größeres Modell einsetzen

Die Rate sinkt mit jeder Generation. Die Ursache bleibt unverändert bestehen.

Auf eigene Daten feinabstimmen

Das verändert Stil und Fachsprache. Wahr werden Aussagen dadurch nicht.

Das Modell sich selbst prüfen lassen

Zwei Läufe desselben Modells teilen denselben blinden Fleck.

In welcher Reihenfolge vorgehen

1. **Aufgabe zuschneiden**
   Alles, was sich rechnen oder nachschlagen lässt, aus dem Modell herausnehmen.
2. **Quellen und Ausgabeform festlegen**
   Woher die Aussagen kommen und welche Form das Ergebnis hat.
3. **Prüfmenge aufbauen, bevor gebaut wird**
   Ohne bekannte Sollwerte lässt sich später nichts messen.

Womit auch danach zu rechnen ist

- Ein richtig belegter Satz kann falsch zusammengefasst sein.
- Die passende Quelle kann übersehen worden sein.
- Eine mehrdeutige Frage kann in der falschen Lesart beantwortet werden.

Halluzinationen werden eingerahmt, nicht beseitigt.
Die brauchbare Frage lautet: Woran fällt ein falscher Wert auf, bevor er wirkt?