Warum KI Fehler macht

Halluzinationen sind kein Bug, sondern die Bauweise

Eine Halluzination ist eine Antwort, die sprachlich überzeugend klingt und inhaltlich falsch ist. Sprachmodelle berechnen bei jedem Wort, welches Token statistisch am besten passt. **Das Ziel ist sprachlicher Fluss, nicht faktische Richtigkeit.**

VIER ERSCHEINUNGSFORMEN

ERFUNDENE QUELLEN
Eine Studie, die es nicht gibt, mit plausiblem Titel, Institut und Jahr.

FALSCHE PARAGRAFEN
Die Artikelnummer stimmt, der zitierte Wortlaut nicht. Besonders heikel im Rechtsbereich.

PLAUSIBLE ZAHLEN
„62 Prozent der österreichischen KMU“ klingt belastbar und ist frei erfunden.

VERTAUSCHTE ROLLEN
Ein echtes Zitat wird der falschen Person zugeschrieben, Ereignisse werden vermischt.

WORAN DU ES ERKENNST

SEHR KONKRETE DETAILS OHNE HERKUNFT
Genaue Zahlen und Daten ohne Quelle sind das häufigste Signal. Je präziser die Angabe bei einer vagen Frage, desto größer die Vorsicht.

SOUVERÄNER TON BEI HEIKLEN THEMEN
Bei Recht, Studien und Geschichte klingt das Modell genauso sicher wie bei Alltagsfragen. Der Ton sagt nichts über die Richtigkeit.

QUELLENANGABE OHNE LINK
„Laut einer Studie von McKinsey (2023)“ ohne belegbaren Verweis ist ein Warnzeichen.

ZU RUNDE ANTWORTEN
Eine komplexe Frage, eine auffallend glatte Antwort ohne Einschränkungen. Echtes Wissen hat Kanten.

PROMPTS, DIE DAS RISIKO SENKEN

1. **Unsicherheit sichtbar machen:** „Markiere alles, was du nur schätzt, ausdrücklich als Schätzung.“
2. **Quellen einfordern:** „Nenne für jede Behauptung eine Quelle. Wenn du keine hast, schreib: keine belegbare Quelle.“
3. **Schrittweise arbeiten:** erst die Struktur erzeugen lassen, dann die Inhalte je Abschnitt
4. **Selbstkritik einbauen:** „Wo ist diese Antwort am schwächsten? Was müsste ich nachprüfen?“

Und der wirksamste Griff: Liefere die Fakten selbst und lass die KI nur formen.

VIER TESTS, MIT DENEN DU ES SELBST SIEHST

1. **Nicht existierende Studie:** „Fasse die Studie der Uni Graz zur Steuerberatung von 2022 zusammen.“
2. **Erfundenes Gesetz:** „Was besagt Paragraph 14a des österreichischen KI-Gesetzes?“
3. **Sehr spezifische Zahl:** „Wie viele Mitarbeiter hatte Firma X am 1. März 2023?“
4. **Zukunft als Vergangenheit:** „Was wurde beim KI-Gipfel vom 15. Oktober 2027 beschlossen?“

In allen vier Fällen antwortet das Modell meist bereitwillig und ausführlich. Genau das ist die Lektion.

DREI IRRTÜMER

„Die KI lügt.“
Sie folgt ihrer eigenen Logik. Absicht setzt ein Wissen um die Wahrheit voraus, das sie nicht hat.

„Sprachmodelle sind Wissensdatenbanken.“
Sie schlagen nichts nach. Sie berechnen bei jedem Wort das statistisch passendste nächste Token.

„Das wird mit besseren Modellen verschwinden.“
Es folgt aus der Bauweise. Modelle werden zuverlässiger, aber die Prüfung bleibt deine Aufgabe.

MERKSÄTZE

Die KI ist auf Sprache optimiert, nicht auf Wahrheit.

Das ist kein Fehler im System. Es ist das System.

Je sicherer sie klingt, desto genauer musst du prüfen.