Rico Eberle · LEARNING NUGGETS

LEARNING NUGGET E-LEARNING & KI 12. SEPTEMBER 2026 (AKTUALISIERT 31. OKTOBER 2026)

Die KI lügt nicht. Sie rät.

Sprachmodelle erfinden Antworten nicht aus Absicht, sondern weil sie das wahrscheinlichste nächste Wort wählen und fürs Raten belohnt werden.

Von Rico Eberle, E-Learning-Experte für Weiterbildung und Gemeinderat in Dübendorf

Video zum Beitrag

Dieses Learning Nugget gibt es als Video mit Transkript (1:19 Min.). QR-Code scannen oder Link öffnen:

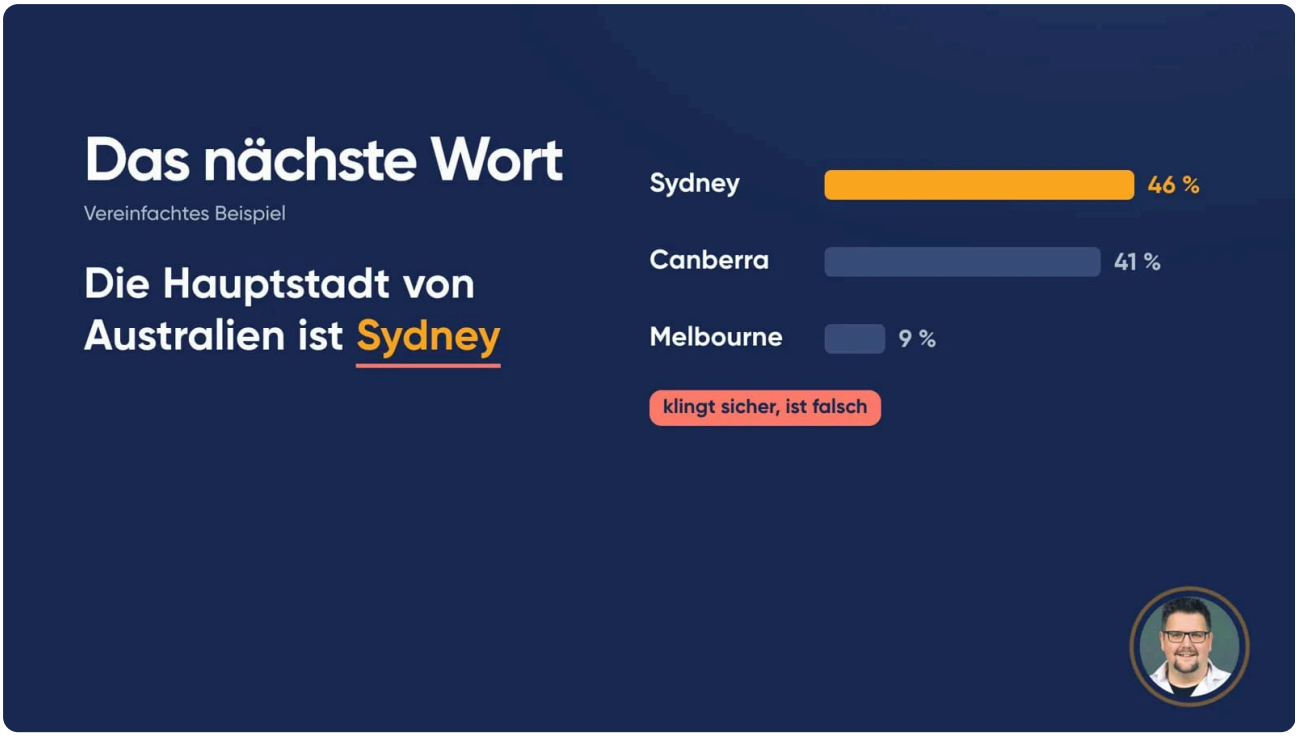
ricoeberle.ch/blog/warum-ki-halluziniert/

2025 hat OpenAI erklärt, warum Sprachmodelle Dinge erfinden. Die kurze Antwort: Falsche Antworten entstehen schon im Training. Und sie halten sich, weil Modelle fürs Raten belohnt werden.

Man nennt das Halluzination. Das Wort klingt nach einer Störung. Tatsächlich folgt sie direkt daraus, wie Sprachmodelle funktionieren und wie sie bewertet werden.

Ein Sprachmodell kennt keine Wahrheit, nur Wahrscheinlichkeiten

Ein Sprachmodell schlägt nicht nach. Es berechnet, vereinfacht gesagt, Wort für Wort, welches Wort am wahrscheinlichsten als nächstes kommt. Weil es mit riesigen Textmengen trainiert wurde, ist das wahrscheinlichste Wort sehr oft auch das richtige.



Das wahrscheinlichste Wort ist nicht immer das richtige: Sydney statt Canberra. Die Prozentwerte sind erfundene Beispielwerte.

Weiss das Modell etwas nicht, ändert sich am Ablauf nichts. Es wählt weiterhin das wahrscheinlichste nächste Wort. Heraus kommt ein Satz, der genauso sicher klingt wie eine korrekte Antwort. Ein Zögern oder ein «vielleicht» kommt nicht automatisch dazu.

Raten bringt Punkte, Schweigen nicht

Die Erklärung von OpenAI erschien zusammen mit einer wissenschaftlichen Arbeit. Sprachmodelle werden ähnlich bewertet wie Lernende bei einer Multiple-Choice-Prüfung. Wer rät, holt manchmal Punkte. Wer nichts ankreuzt, holt sicher keine.

Warum raten?

Raten	manchmal Punkte
Leer lassen	sicher 0 Punkte

Belohnt die Bewertung das Raten und bestraft sie Zurückhaltung, lernt ein Modell: Raten lohnt sich. Die Halluzination ist also auch eine Folge davon, wie Modelle gemessen und verbessert werden.

Der sichere Ton sagt nichts über die Richtigkeit

Das ist die wichtigste Einsicht für Lernende. Eine KI-Antwort ist ein Entwurf, kein Nachschlagewerk.

Drei Prüffragen helfen im Alltag:

1. **Finde ich das in einer zweiten, unabhängigen Quelle?** Nicht in einer zweiten KI, sondern in einer Quelle, die selbst belegt ist.
2. **Lassen sich Namen, Zahlen und Zitate konkret überprüfen?** Diese Details sind besonders anfällig, weil sie plausibel klingen müssen, aber exakt stimmen sollten.
3. **Was passiert, wenn ich dieselbe Frage anders stelle?** Ändert sich die Antwort bei einer leicht anderen Formulierung, ist das ein Warnsignal.

Eine Übung für die nächste Lektion

Lass die Lernenden in Gruppen eine KI zu einem Thema befragen, das sie bereits gut kennen. Jede Gruppe sucht mit den drei Prüffragen eine Stelle, die nicht stimmt, und belegt den Fehler mit einer Quelle.

Im Plenum zeigt sich meist schnell, wie überzeugend falsche Antworten wirken können. So wird aus Misstrauen eine Technik, die man üben kann.

Die KI ist kein Lexikon, sondern ein sehr guter Rater.

Häufige Fragen

Warum halluziniert KI?

Ein Sprachmodell berechnet Wort für Wort, welches Wort am wahrscheinlichsten als nächstes kommt. Weiss es etwas nicht, wählt es trotzdem das wahrscheinlichste Wort, und das klingt genauso sicher wie eine richtige Antwort.

Warum sagt die KI nicht einfach «Ich weiss es nicht»?

Laut einer Erklärung von OpenAI (2025) werden Sprachmodelle ähnlich bewertet wie Lernende bei einer Multiple-Choice-Prüfung: Raten bringt manchmal Punkte, nichts ankreuzen sicher keine. So lernt ein Modell, dass Raten sich lohnt.

Wie erkennt man erfundene KI-Antworten?

Drei Prüffragen helfen: Finde ich das in einer zweiten, unabhängigen Quelle? Lassen sich Namen, Zahlen und Zitate konkret überprüfen? Ändert sich die Antwort, wenn ich dieselbe Frage anders stelle?

Quellen

1. OpenAI: Why language models hallucinate (2025)
   https://openai.com/index/why-language-models-hallucinate/
2. Kalai et al.: Why Language Models Hallucinate (2025)
   https://arxiv.org/abs/2509.04664

ÜBER MICH

Rico Eberle

Ich arbeite im E-Learning, bin Betriebsökonom FH und Gemeinderat in Dübendorf und präsidiere den Stiftungsrat der Stiftung Weiterbildungskurse Dübendorf (WBK). Die Learning Nuggets mache ich allein: Thema, Recherche, Text und Video. Fragen und Widerspruch sind willkommen.

Kontakt: ricoeberle.ch · **LinkedIn:** linkedin.com/in/rico-eberle

LIZENZ

Text, Transkript und Video von Rico Eberle stehen unter der Lizenz CC BY 4.0 (ausgenommen Musik und Soundeffekte des Videos). Du darfst die Inhalte teilen und bearbeiten, auch im Unterricht, wenn du die Quelle nennst.
creativecommons.org/licenses/by/4.0/deed.de

ZITIERVORSCHLAG

«Die KI lügt nicht. Sie rät.», Rico Eberle, ricoeberle.ch, 2026, https://ricoeberle.ch/blog/warum-ki-halluziniert/, CC BY 4.0