

LEARNING NUGGET E-LEARNING & KI 12. SEPTEMBER 2026

Die KI lügt nicht. Sie rät.

Warum Sprachmodelle Dinge erfinden, warum sie dabei so sicher klingen und mit welchen drei Prüffragen Lernende erfundene Antworten erkennen.

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

RICO EBERLE · LEARNING NUGGET

Die KI lügt nicht. Sie rät.

Und klingt dabei immer sicher.



Video zum Beitrag

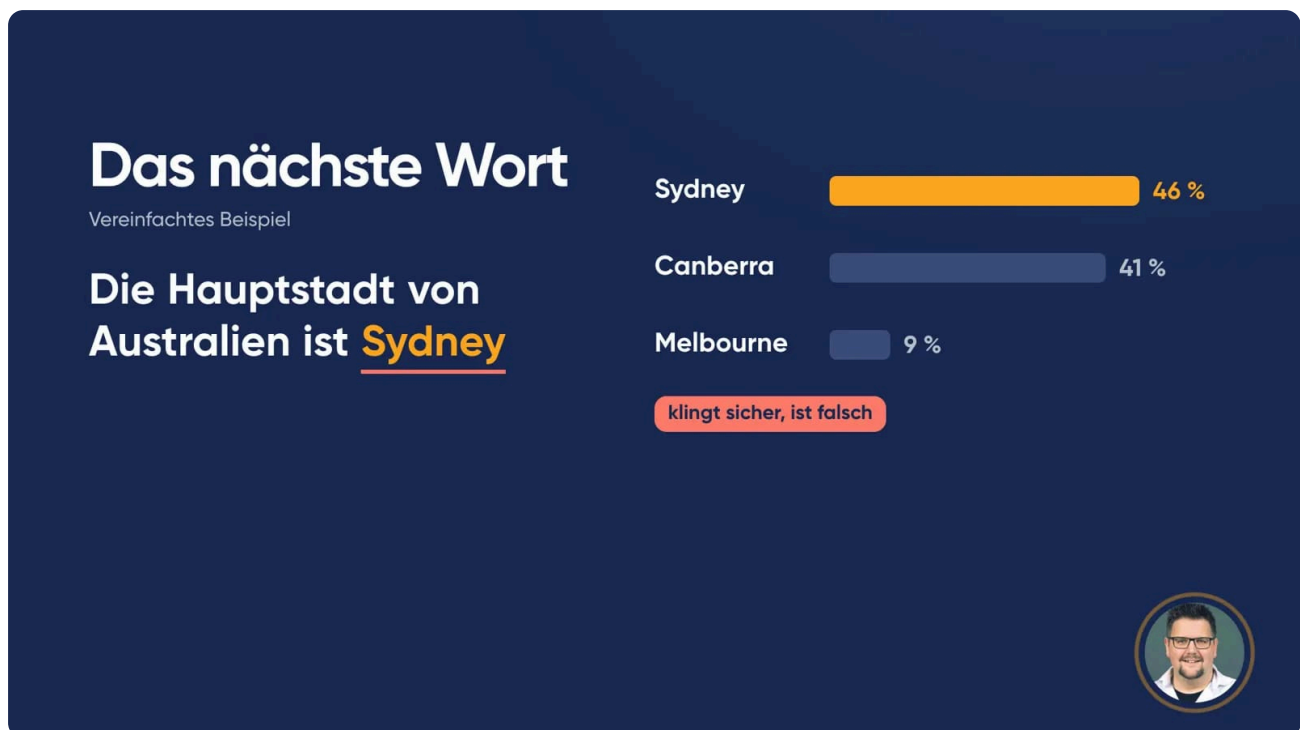
Dieses Learning Nugget gibt es als Video mit Transkript (1:19 Min.). QR-Code scannen oder Link öffnen:

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

Wer mit KI-Werkzeugen arbeitet, kennt den Moment: Die Antwort ist flüssig formuliert, gut strukturiert und klingt kompetent. Und trotzdem stimmt sie nicht. Eine Quelle, die es nicht gibt. Eine Jahreszahl, die knapp daneben liegt. Ein Zitat, das nie gesagt wurde. Man nennt das Halluzination. Das Wort klingt nach einer Störung. Tatsächlich ist es eine direkte Folge davon, wie Sprachmodelle funktionieren.

Ein Sprachmodell kennt keine Wahrheit, nur Wahrscheinlichkeiten

Ein Sprachmodell schlägt nicht nach. Es berechnet, vereinfacht gesagt, welches Wort am wahrscheinlichsten als nächstes kommt, und das Wort für Wort. 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.

Weiss das Modell etwas nicht, ändert sich an diesem Ablauf nichts. Es wählt weiterhin das wahrscheinlichste nächste Wort. Das Ergebnis ist ein Satz, der genauso sicher klingt wie eine korrekte Antwort. Ein Zögern, ein «vielleicht» oder einen Hinweis auf Unsicherheit gibt es dabei nicht automatisch.

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

OpenAI hat 2025 eine Erklärung dazu veröffentlicht, begleitet von einer wissenschaftlichen Arbeit. Der Kern: 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

Also lernt das Modell: Raten lohnt sich.

Wie bei einer Multiple-Choice-Prüfung bringt Raten im Schnitt mehr Punkte als Schweigen.

Wenn die Bewertung Raten belohnt und Zurückhaltung bestraft, lernt ein Modell genau das: Raten lohnt sich. Die Halluzination ist also auch ein Ergebnis davon, wie Modelle gemessen und verbessert werden.

Was das für den Unterricht heisst

Die wichtigste Einsicht für Lernende lautet: **Der sichere Ton sagt nichts über die Richtigkeit.** Eine KI-Antwort ist ein Entwurf, kein Nachschlagewerk.

Drei Prüffragen

Sicherer Ton ≠ richtige Antwort

- 1 Finde ich das in einer zweiten, unabhängigen Quelle?
- 2 Lassen sich Namen, Zahlen und Zitate überprüfen?
- 3 Was passiert, wenn ich dieselbe Frage anders stelle?



Drei Prüffragen für jede KI-Antwort.

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?** Gerade 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 nach einer Stelle, die nicht stimmt, und belegt den Fehler mit einer Quelle. Die Auswertung im Plenum zeigt meist schnell, wie überzeugend falsche Antworten wirken können. Und sie macht aus dem Misstrauen eine Technik, die man üben kann.

Die KI ist kein Lexikon, sondern ein sehr guter Rater. Wer das weiss, kann sie sinnvoll nutzen.

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 DEN AUTOR

Rico Eberle

Rico Eberle ist E-Learning-Experte, Betriebsökonom FH und Gemeinderat in Dübendorf. Er präsidiert den Stiftungsrat der Stiftung Weiterbildungskurse Dübendorf (WBK). In den Learning Nuggets erklärt er Forschung zu Lernen, KI und digitaler Souveränität kurz und mit Quellen.

Kontakt: ricoeberle.ch · LinkedIn: [linkedin.com/in/rico-eberle](https://www.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