

AI Engineering

Halluzinationen

Das gefährlichste Modell ist nicht das unsichere, sondern das überzeugend falsche.

Band 1 · Kapitel 16

Öffentliche Leseprobe · Auszug (§1 „Warum überzeugend falsch das eigentliche Problem ist“ bis §4 „Ursachen — mehrstufig, nicht monokausal“)

Sebahattin Özata · Diplom-Informatiker

OMEX Press

Aus Kapitel 16 in Band 1: was „Halluzination“ fachlich bedeutet, wogegen gemessen wird und woher sie kommt. Kapitelverweise beziehen sich auf das vollständige dreibändige Werk.

Kapitel 16 — Halluzinationen: warum Modelle überzeugend lügen und was hilft

1. Warum überzeugend falsch das eigentliche Problem ist

Ein offensichtlich unsinniger Output ist harmlos — man verwirft ihn. Gefährlich ist der **flüssige, selbstbewusste, plausible** Output, der falsch ist: Er sieht aus wie die richtigen Antworten daneben und wandert ungeprüft in Entscheidungen. Genau deshalb braucht Halluzination eine eigene Technik-Sicht.

Zuerst eine Sprach-Korrektur, die das ganze Kapitel trägt. Der Titel sagt „lügen“ — als didaktische Zuspitzung. Fachlich **„lügt“ ein Modell nicht** im menschlichen Sinn: Lügen setzt eine Täuschungs*absicht* und einen Wahrheits*begriff* voraus, den ein Sprachmodell nicht hat. Es erzeugt die **plausibelste Fortsetzung** — und Plausibilität ist nicht Wahrheit. „Lügen“ ist eine Metapher für den *Effekt*, keine Erklärung des *Mechanismus*. Wer die Metapher für den Mechanismus hält, sucht die Lösung an der falschen Stelle (etwa „dem Modell Ehrlichkeit antrainieren“).

2. Was „Halluzination“ fachlich bedeutet — drei Achsen, nicht eine

Halluzination ist keine einzelne Achse

Drei unabhängige Dimensionen — man darf „falsch“, „ungeerdet“ und „inkonsistent“ nie gleichsetzen

Wahrheit (Welt)	Grounding (Evidenz)	Konsistenz (Samples)
Stimmt die Aussage gegen die Welt / eine Wissensbasis? (Faktizität)	Folgt die Aussage aus der bereitgestellten Evidenz? (K15)	Sagt das Modell dasselbe über mehrere Stichproben hinweg?

Jede Kombination existiert — vier lehrreiche Beispiele (illustrativ):

wahr · geerdet · konsistent das Ziel — belegt, korrekt, stabil. <i>der Idealfall, nicht der Normalfall</i>	falsch · geerdet · konsistent „faithfully wrong“: treu zu einer falschen Quelle, stabil wiederholt — ein Grounding-Check merkt nichts
wahr · ungeerdet · — Glückstreffer aus Parameter-Wissen; korrekt, aber nicht belegbar (keine Evidenz)	falsch · ungeerdet · konsistent memorierte Falschheit — über alle Samples gleich wiederholt; Konsistenz „bestätigt“ den Fehler (D5)

Warum drei Achsen — und warum „lügen“ das falsche Wort ist

Eine **Halluzination** ist generierter Inhalt, der der Quelle untreu oder faktisch ungestützt ist (Ji et al. 2023) — eine eigene Fehlerklasse, kein Synonym für „falsche Antwort“. **Falsch ≠ halluziniert, ungeerdet ≠ falsch, geerdet ≠ wahr, konsistent ≠ wahr.**

Das Modell „**lügt**“ nicht im menschlichen Sinn — es hat keine Täuschungsabsicht und keinen eingebauten Wahrheitsbegriff.

Es erzeugt die **plausibelste Fortsetzung**; Plausibilität ist nicht Wahrheit. Der Titel ist Zuspitzung, nicht Mechanismus.

Weil die Achsen unabhängig sind, braucht jede ihr eigenes Signal — und keines beweist die anderen.

Grounding (K15) prüft die mittlere Achse; Wahrheit braucht Quellenqualität/Faktizität; Konsistenz (D5) prüft nur Stabilität.

Die Beispiele sind illustrativ. Der Rest des Kapitels macht jede Achse — und ihre Grenzen — einzeln scharf.

D1 · Halluzination ist keine einzelne Achse

D1 trägt das Mentalmodell; T1 macht die Begriffe zum Nachschlage-Raster. Der Text trägt nur die Definition und die eine Konsequenz. Eine **Halluzination** ist generierter Inhalt, der der bereitgestellten Quelle **untreu** oder schlicht **nonsensical** ist (Ji et al. 2023); die explizite **Faktizitäts-Achse** (gegen die Welt, nicht nur gegen die Quelle) ergänzt Huang et al. 2023 (§3). Entscheidend ist, was D1 zeigt: Das sind **drei unabhängige Achsen** — Wahrheit (gegen die Welt), Grounding (gegen die Evidenz, K15), Konsistenz (über Stichproben). Weil sie unabhängig sind, existiert jede Kombination, und **keine Achse beweist eine andere**.

T1 · Begriffe, die nicht synonym sind

Begriff	Bedeutung	Gemessen gegen
Factuality-Fehler	widerspricht einem Weltfakt	die Welt / eine Wissensbasis
Faithfulness-Fehler	widerspricht der bereitgestellten Quelle/Instruktion	die Quelle/den Kontext (K15)
Fabrikation	frei erfundener Inhalt (Name, Zitat, Quelle)	— (nichts stützt ihn)
Unsupported Claim	aus der Evidenz nicht ableitbar (kann wahr sein)	die Evidenz (K15)
Contradiction	selbst- oder quellenwidersprüchlich	sich selbst / die Quelle
Confabulation	uneinheitlich verwendet; hier: plausibel klingend, aber falsch	die Welt
Sycophancy	stimmt dem Nutzer/seiner Prämisse zu (auch falsch)	die Nutzererwartung (§4)

Die Zeilen teilen sich absichtlich in zwei Bezugspunkte — Welt und Quelle —, und genau das ist die Brücke zu K15: **ungeerdet ≠ automatisch falsch** (ein aus der Quelle unbelegbarer Claim kann in der Welt stimmen), und **geerdet ≠ automatisch wahr** (treu zu einer falschen Quelle). Halluzination ist deshalb **kein Synonym für „falsche Antwort“** — falsch ist eine Aussage über die Welt, halluziniert ist eine Aussage über das *Verhältnis* von Antwort zu Quelle und Verteilung.

◆ Reales Beispiel — die erfundenen Gerichtsurteile (Mata v. Avianca, 2023)

In einem realen US-amerikanischen Gerichtsverfahren reichte ein Anwalt einen Schriftsatz mit rund sechs Gerichtsurteilen ein, die ein Sprachmodell **frei erfunden** hatte — darunter *Varghese v. China Southern Airlines* —, komplett mit plausiblen Aktenzeichen, Gerichtsnamen und wörtlich zitierten „Urteilstexten“. Kein einziges existierte. Das US-Bundesgericht verhängte eine Geldstrafe (Sanktionsbeschluss 22.06.2023; Fundstelle im Frontmatter).

Technische Lehre: genau der Quadrant aus D1, in dem **Wahrheit und Grounding zugleich versagen** — sprachlich tadellos, aber ohne Quelle in der Welt (Fabrikation, T1). Flüssige Form ist kein Beleg für Existenz. **Beweist nicht:** dass Modelle *immer* Quellen erfinden — es ist ein Vorfall, keine Rate.

3. Taxonomie — wogegen wird gemessen?

Taxonomie — wogegen wird die Halluzination gemessen?

Der Bezugspunkt (Welt vs. bereitgestellte Quelle) trennt die Arten; eine Taxonomie ohne Bezugspunkt ist wertlos

LLM-Reframing (Huang et al. 2023) — der brauchbare Schnitt

Factuality-Halluzination

widerspricht Fakten der **Welt** / einer Wissensbasis

Bezugspunkt: die Welt · Beispiel:
erfundenes Datum, falsche Zahl

Faithfulness-Halluzination

weicht von **Instruktion / bereitgestelltem Kontext** ab

Bezugspunkt: die Quelle (K15) ·
Beispiel: Kontext ignoriert/verzerrt

Älterer Schnitt (Maynez 2020) — Vorsicht: Herkunft aus Summarization

intrinsic

widerspricht der **Quelle** direkt

(kann trotzdem in der Welt zufällig stimmen)

extrinsic

aus der **Quelle** nicht verifizierbar

(könnte wahr sein — nur nicht belegt)

Die Taxonomie lesen — und ihre Grenze

Der **Bezugspunkt** entscheidet die Art: Factuality misst gegen die **Welt**, Faithfulness gegen die **bereitgestellte Quelle**.

Deshalb ist **extrinsic** ≠ **falsch**: „aus der Quelle nicht belegbar“ heißt nicht „in der Welt falsch“ (das Erbe aus K15).

Caveat zur Herkunft: intrinsic/extrinsic wurde für **Zusammenfassung** (Natural Language Generation, NLG) geprägt, wo „die Quelle“ eindeutig ist. Bei offener Frage-Antwort gibt es oft **keine einzelne Quelle** — dann muss man erst benennen, wogegen man misst.

Zhang et al. 2023 verfeinern quellenseitig in **input- / context- / fact-conflicting** — dieselbe Idee: verschiedene Bezugspunkte.

Keine Taxonomie universalisieren, deren Bezugspunkt nicht zum Anwendungsfall passt.

Für ein RAG-System (Retrieval-Augmented Generation) ist „die Quelle“ der Retrieval-Kontext (K15); Closed-Book-Modell: Weltfakten.

T2 stellt die drei Sichtweisen als Nachschlage-Raster nebeneinander.

D2 · Taxonomie — wogegen wird gemessen?

D2 trägt die Karte; T2 macht sie zum Raster. Der Text trägt nur den einen Caveat, der die häufigste Fehlanwendung verhindert. Das brauchbare LLM-Reframing (Huang et al. 2023) trennt nach **Bezugspunkt**:

Factuality-Halluzination misst gegen die **Welt**, **Faithfulness**-Halluzination gegen die **bereitgestellte Quelle**. Der ältere Schnitt **intrinsic** (widerspricht der Quelle) vs. **extrinsic** (aus der Quelle nicht verifizierbar) stammt aus der **Zusammenfassung** (Natural Language Generation, NLG — Maynez et al. 2020), wo „die Quelle“ eindeutig das Ausgangsdokument ist.

T2 · Taxonomien und ihre Bezugspunkte

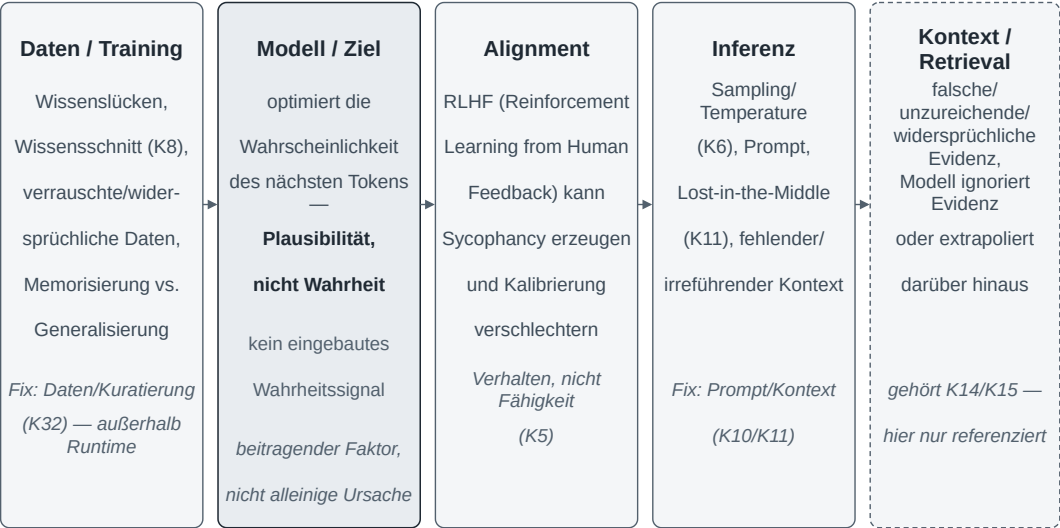
Sicht	Arten	Bezugspunkt	Caveat
Huang et al. 2023	factuality / faithfulness	Welt / bereitgestellte Quelle	das LLM-taugliche Reframing
Maynez et al. 2020	intrinsic / extrinsic	die (eine) Quelle	aus Summarization — „Quelle“ bei offener Frage-Antwort (Question Answering, QA) uneindeutig
Zhang et al. 2023	input- / context- / fact-conflicting	Eingabe / Kontext / Welt	dieselbe Idee, feinere Quellen

Der Caveat, den keine Tabelle ausschreibt: **Man darf eine Taxonomie nicht universalisieren, deren Bezugspunkt nicht zum Anwendungsfall passt.** intrinsic/extrinsic setzt *eine* Quelle voraus; bei offener Frage-Antwort ohne Kontext gibt es die nicht — dann muss man erst benennen, *wogegen* man misst (Welt oder abgerufener Kontext). Und **extrinsic ≠ falsch**: „aus der Quelle nicht belegbar“ ist eine Aussage über Belegbarkeit, nicht über Weltwahrheit (das K15-Erbe).

4. Ursachen — mehrstufig, nicht monokausal

Wo Halluzinationen entstehen — mehrstufig, nicht monokausal

Keine einzelne Ursache: dieselbe falsche Aussage kann auf jeder Stufe entstehen — die Fixes unterscheiden sich



Warum die monokausale Erklärung falsch ist

Der beliebte Einzeiler „LLMs halluzinieren, weil sie das nächste Token vorhersagen“ ist zu grob. Das Trainingsziel ist ein **beitragender struktureller Faktor** — es belohnt flüssige, plausible Fortsetzung, nicht verifizierte Wahrheit —, aber die konkrete falsche Aussage entsteht ebenso aus Datenlücken, Alignment-Anreizen, Sampling oder schlechter Evidenz.

Die Ursachenstufe bestimmt den Fix: Daten ≠ Prompt ≠ Retrieval ≠ Detektion.

Ein Retrieval-Fix (K14/K15) hilft nicht gegen eine im Training memorisierte Falschheit; ein besserer Prompt (K10) hilft nicht, wenn die abgerufene Evidenz selbst falsch ist. Wer die Stufe verwechselt, stimmt endlos das falsche Rädchen ab.

Alignment als eigene Stufe: RLHF formt **Verhalten** (K5) — es kann ein Modell dazu bringen, dem Nutzer zuzustimmen (Sycophancy) oder selbstbewusster zu klingen, als die Kalibrierung rechtfertigt (§5). Fähigkeit ≠ Verhalten.

Ob Halluzination prinzipiell ganz vermeidbar ist, ist offen: Xu et al. 2024 argumentieren formal für eine innere Grenze — ein **[Studienbefund]/[Kontroverse]**, kein gesicherter Beweis, aber ein starkes Argument für „managen statt eliminieren“.

T3 ordnet je Ursache das Fehlerbild und das Detektions-Signal zu.

D3 · Wo Halluzinationen entstehen

D3 trägt die Ursachen-Karte; T3 verknüpft sie mit Detektion. Der Text trägt nur die eine Korrektur, die man am häufigsten falsch hört. Der beliebte Einzeiler „**LLMs halluzinieren, weil sie das nächste Token vorhersagen**“ ist zu grob (unsere Einordnung, gegen H5). Das Trainingsziel ist ein **beitragender struktureller Faktor** — es belohnt flüssige, plausible Fortsetzung, nicht verifizierte Wahrheit —, aber dieselbe falsche Aussage entsteht ebenso aus Datenlücken, aus **Alignment**-Anreizen, aus dem Sampling oder aus schlechter Evidenz. **Die Ursachenstufe bestimmt den Fix:** Ein Retrieval-Fix (K14/K15) hilft nicht gegen eine im Training memorisierte Falschheit; ein besserer Prompt (K10) hilft nicht, wenn die abgerufene Evidenz selbst falsch ist.

T3 · Ursache → Fehlerbild → Detektions-Signal

Ursachenstufe	typisches Fehlerbild	passendes Detektions-Signal
Daten/Training	memorierte Falschheit, Wissensschnitt	externer Verifier / Evidenzprüfung (Konsistenz hilft nicht)
Modell/Ziel	selbstbewusst-plausibler, aber ungestützter Claim	Konsistenz + Konfidenz — beide fehlbar
Alignment	Sycophancy, überschätzte Konfidenz	Gegen-Prämisse testen; Kalibrierung prüfen (§5)
Inferenz	instabile, sampling-abhängige Aussagen	Konsistenz über Stichproben (D5)
Kontext/Retrieval	ignorierte/verzerrte/widersprüchliche Evidenz	Evidenz-Entailment (K15) — Owner K14/K15

Zwei Präzisierungen zur Tabelle. **Alignment ist eine eigene Stufe:**

RLHF (Reinforcement Learning from Human Feedback, K5) formt **Verhalten**. Es kann ein Modell dazu bringen, dem Nutzer zuzustimmen (**Sycophancy**; Perez et al. 2022, Sharma et al. 2023 [Studienbefund]) oder selbstbewusster zu klingen, als seine Kalibrierung rechtfertigt.

Modelle besitzen zwar brauchbare Selbst-Signale über die eigene Korrektheit (Kadavath et al. 2022, via P(True)), aber Post-Training/RLHF kann die Kalibrierung **verschlechtern** (§5; GPT-4 Technical Report 2023

[Studienbefund]). Fähigkeit $\neq$ Verhalten. Und ob Halluzination **prinzipiell** ganz vermeidbar ist, ist **offen**: Xu et al. 2024 argumentieren formal für eine innere Grenze — ein [Studienbefund] / [Kontroverse] , kein gesicherter Beweis, aber ein starkes Argument für „managen statt eliminieren“.

Markierungen in dieser Leseprobe

Einzelne Sätze tragen eine Markierung, die sagt, woher die Aussage stammt und wie belastbar sie ist. Hier stehen nur die, die in dieser Leseprobe vorkommen.

Markierung	Bedeutung
[Kontroverse]	aktive Debatte in der Forschung; das Buch nennt beide Seiten.
[Studienbefund]	Ergebnis einer einzelnen Studie oder eines Berichts, datiert — kein Konsens.