HANDOUT KRITISCHES THEMA KI-HALLUZINATIONEN

KI-Halluzinationen

Handout-Materialien zum Workshop — Verstehen, erkennen, vermeiden

5 Materialien · Erstellt: März 2026 · Version 1.0

Was sind KI-Halluzinationen?

Wenn KI Fakten erfindet — und dabei überzeugend klingt

> Ein Sprachmodell hat kein Wissen — es hat Muster. Es wählt immer das statistisch plausibelste nächste Wort. Das funktioniert meistens. Aber wenn das häufigste Muster falsch ist, klingt die Antwort trotzdem überzeugend.

WIE SPRACHMODELLE FUNKTIONIEREN

Sprachmodelle wie ChatGPT, Copilot oder Gemini haben aus Milliarden von Texten gelernt, welche Wörter häufig zusammen vorkommen. Sie speichern keine Fakten in einer Datenbank — sie lernen statistische Muster. Für jedes neue Wort berechnet das Modell: "Was folgt hier am wahrscheinlichsten?"

> **Das Problem:** Das häufigste Muster ist nicht immer das richtige. "Die Hauptstadt von Australien ist Sydney" — klingt plausibel, ist falsch. Sydney ist häufiger mit Australien verknüpft als Canberra. Das Modell wählt Sydney. Und sagt es mit Selbstsicherheit.

KONKRETES BEISPIEL: HARVARD-FAKE-STUDIE

KI	"Harvard Business Review, 2021. Studie zeigt: 73 % der Führungskräfte bevorzugen asynchrone Kommunikation in hybriden Teams."
FAKT	Diese Studie existiert nicht. Weder Titel noch Zahl noch Autoren lassen sich verifizieren. Das Modell hat ein plausibel klingendes Zitat konstruiert.

WEITERES BEISPIEL: PRODUKTIVITÄTSZAHL

KI	"Laut McKinsey-Studie 2023 steigert KI-Einsatz die Produktivität um 47 %."
FAKT	Es gibt McKinsey-Studien zu KI-Produktivität. Aber diese exakte Zahl aus diesem Erscheinungsjahr? Nicht verifizierbar. Die Zahl klingt präzise genug um glaubwürdig zu sein.

> *„ Sprachmodelle erzeugen Text auf Basis von Wahrscheinlichkeiten — nicht auf Basis von Wissen. Selbstsicherheit in der Antwort ist kein Indikator für Korrektheit. “*

Wie häufig passiert das?

Zahlen zur Halluzinationsrate in der Praxis

Halluzinationen sind kein Randphänomen. Sie treten regelmäßig auf — und meistens dort, wo sie am meisten schaden: bei konkreten Zahlen, Studien und Quellenangaben.

≤ 20 %

Halluzinationsrate

Bei faktischen Abfragen halluzinieren führende Modelle in Tests bei bis zu 20 % der Fälle mindestens einmal

76 %

Nicht geprüft

Nutzer übernehmen KI-Aussagen ohne Gegenkontrolle in rund 76 % der Fälle — laut Stanford-Studie

3

Fragen reichen

Drei gezielte Rückfragen reichen oft aus, um eine Halluzination aufzudecken — wenn man weiß, wie

Quellen: vgl. Stanford HAI, 2024; Microsoft AI Research, 2023 (Richtwerte; eigene Verifizierung empfohlen)

Besonders riskant bei: konkreten Prozentzahlen, Studientiteln und Autoren, historischen Fakten, rechtlichen Details, Medikamentendosierungen, Unternehmenskennzahlen

4 Strategien gegen Halluzinationen

Konkrete Maßnahmen für den täglichen Umgang mit KI

Du kannst Halluzinationen nicht vollständig verhindern — aber du kannst sie systematisch aufdecken und ihnen vorbeugen. Diese vier Strategien reichen für den Alltag.

01

Kleinere Schritte statt Mammutaufgaben

Je komplexer der Auftrag, desto mehr Raum für Fehler. Teile große Anfragen in kleine, prüfbare Einheiten auf. Eine Teilaufgabe lässt sich verifizieren — eine 20-seitige Analyse nicht mehr.

✓ Statt: "Erstelle eine vollständige Marktanalyse" → Schritt für Schritt vorgehen, jeden Abschnitt einzeln prüfen

WANN NICHT?

Bei kurzen, klar abgrenzbaren Aufgaben (eine E-Mail umformulieren, einen Begriff erklären) ist die Zerteilung unnötig — direkter Prompt reicht.

02

Unsicherheit explizit einfordern

Sprachmodelle signalisieren von sich aus selten Zweifel. Frag aktiv: "Wie sicher bist du?" oder "Was weißt du nicht?" Das Modell antwortet dann ehrlicher über seine Wissensgrenzen.

✓ Prompt-Ergänzung: "Markiere explizit, wo du dir unsicher bist."

WANN NICHT?

Nicht nötig bei kreativen Aufgaben oder Brainstormings, wo Faktentreue keine Rolle spielt — nur einsetzen wenn die Aussagen weiterverwendet werden.

03

Quellen anfordern — und tatsächlich prüfen

Lass dir Quellen nennen. Aber nicht als Beruhigungsmittel — sondern als Startpunkt für eigene Recherche. Eine genannte Quelle ist keine verifizierte Quelle.

✓ Regel: Jede konkrete Zahl und jede Studie wird eigenständig nachgeschlagen — immer.

WANN NICHT?

Bei internen Analysen, Brainstormings oder Textentwürfen, die kein Faktenmaterial enthalten — Quellenprüfung nur dort wo tatsächlich Fakten stehen.

04

Cross-Check bei kritischen Fakten

Bei Zahlen, Daten oder Aussagen die in Berichte, Präsentationen oder E-Mails fließen: zweite Quelle. Nicht weil das Modell lügt — sondern weil es nicht weiß, wann es irrt.

✓ Faustregel: Was öffentlich wird oder Entscheidungen beeinflusst, wird verifiziert.

WANN NICHT?

Für interne Ideation, persönliche Notizen oder kreative Entwürfe ist ein Cross-Check unverhältnismäßig — Aufwand nur dann investieren, wenn das Ergebnis nach außen geht.

„ Nicht jede KI-Aussage braucht einen Faktencheck. Aber alles, was zitiert, veröffentlicht oder zur Entscheidungsgrundlage wird: ja. “

Wie gut ist mein Ergebnis?

Formal korrekt ≠ wirklich gut — woran du den Unterschied erkennst

Viele Nutzer halten eine KI-Antwort für „gut", weil sie flüssig und vollständig klingt. Das ist kein Qualitätsmerkmal. Ein wirklich gutes Ergebnis zeigt erkennbare Wissensgrenzen — und macht Unsicherheiten sichtbar statt zu kaschieren.

VORHER / NACHHER

FORMAL KORREKT Prompt: "Welche Studien belegen den ROI von KI-Trainings in Unternehmen?" → Antwort: Fließende Aufzählung mit Prozentzahlen, Studientiteln und Autorennamen. Selbstsicher, gut lesbar.

PROBLEM Drei der vier genannten Studien existieren nicht. Die Zahlen sind plausibel konstruiert — aber nicht verifizierbar. Du hättest es nicht bemerkt.

WIRKLICH GUT Prompt: "Welche Studien belegen den ROI von KI-Trainings? Sag mir explizit, wo du dir unsicher bist und nenne nur Quellen die du wirklich kennst." → Antwort: Zwei verifizierbare Quellen, klarer Hinweis auf die Lücken in der Forschungslage.

ERGEBNIS Weniger, aber verlässlich. Du weißt was du verwenden kannst — und was nicht.

DREI SIGNALE FÜR WIRKLICH GUTE ERGEBNISSE

Das Modell benennt, was es nicht weiß. Wenn eine Antwort auf alles eine Antwort hat — ohne eine einzige Einschränkung — ist das kein gutes Zeichen. Seriöse Antworten haben Lücken.

Quellen lassen sich tatsächlich finden. Ein kurzer Gegencheck (Google Scholar, offizielle Websites) bestätigt Titel und Autoren — das ist das Minimum vor jeder Weitergabe.

Das Ergebnis klingt nicht zu perfekt. Wenn eine Antwort exakt das liefert was du hören wolltest — runde Zahlen, passende Studien, keine Einschränkungen — dann nochmal genauer hinschauen.

„ Die KI liefert Wörter. Wer prüft, liefert Wahrheit. “

Prompt-Erweiterungen

Fertige Formulierungen — einfach kopieren und einfügen

Diese Formulierungen kannst du an jeden bestehenden Prompt anhängen. Sie kosten dich 5 Sekunden — und reduzieren das Halluzinationsrisiko deutlich.

UNSICHERHEIT EINFORDERN

STRATEGIE 2 · UNSICHERHEIT MARKIEREN

```
Markiere explizit, wo du dir unsicher bist. Kennzeichne hypothetische Aussagen ausdrücklich als solche.
```

STRATEGIE 2 · WISSENSGRENZEN BENENNEN

```
Nenne mir explizit, was du zu diesem Thema nicht weißt oder nicht sicher weißt.
```

QUELLEN SICHERN

STRATEGIE 3 · VERIFIZIERBARE QUELLEN

```
Gib nur Quellen an, die du verifizieren kannst. Keine Quellen erfinden — lieber keine Quelle als eine falsche.
```

STRATEGIE 3 · QUELLEN MIT DETAILS

```
Falls du Studien oder Quellen nennst: Gib Titel, Autor und Erscheinungsjahr an. Wenn du dir bei einer Quelle nicht sicher bist, sag es explizit.
```

SCHRITTWEISE VORGEHEN

STRATEGIE 1 · SCHRITTWEISE AUSFÜHRUNG

```
Zerlege diese Aufgabe in 3 Schritte und zeige mir erst Schritt 1. Warte auf meine Rückmeldung, bevor du weitermachst.
```

KOMBINATION · VOLLSTÄNDIGE SICHERHEITSFORMEL

```
Beantworte das schrittweise. Markiere, wo du dir unsicher bist. Nenne nur Quellen, die du wirklich kennst — lieber keine als eine erfundene.
```

Checkliste: Bevor du eine KI-Aussage weitergibst

5 Punkte — 60 Sekunden — kein Reputationsrisiko

> Nicht jede KI-Ausgabe braucht eine Vollprüfung. Aber bevor du eine Aussage zitierst, veröffentlichst oder als Grundlage für Entscheidungen verwendest — nimm dir 60 Sekunden für diese 5 Punkte.

- [ ] **Enthält die Aussage konkrete Zahlen oder Prozentwerte?**
  Zahlen sind halluzinationsanfällig. Spezifische Werte wie "47 %" oder "73 %" wirken präzise — wurden aber oft vom Modell generiert, nicht aus einer Quelle entnommen.

- [ ] **Werden Studien, Berichte oder Autoren genannt?**
  Studientitel, Autorennamen und Erscheinungsjahre werden besonders häufig halluziniert. Jede genannte Quelle: einmal kurz googeln oder direkt nachschlagen.

- [ ] **Klingt die Aussage "zu perfekt"?**
  Wenn eine Antwort exakt das sagt, was du hören wolltest — mit einer passenden Zahl und einer klingenden Quelle — ist Vorsicht geboten. Das Modell optimiert auf Plausibilität, nicht auf Wahrheit.

- [ ] **Habe ich die KI nach ihrer Unsicherheit gefragt?**
  Wenn nicht: nachfragen. "Wie sicher bist du bei dieser Aussage?" und "Was weißt du dazu nicht?" liefern oft überraschend ehrliche Antworten — wenn man fragt.

- [ ] **Geht die Aussage in eine Entscheidung, einen Bericht oder eine E-Mail?**
  Wenn ja: verifizieren. Nicht weil KI unzuverlässig ist — sondern weil die Verantwortung für die Weitergabe bei dir liegt. Das Modell trägt keine Konsequenzen. Du schon.

„ Die KI liefert Wörter; du lieferst die Wahrheit. “