heise KI PRO

KI-Halluzinationen

Ursachen und wie man sie vermeiden kann

heise KI PRO

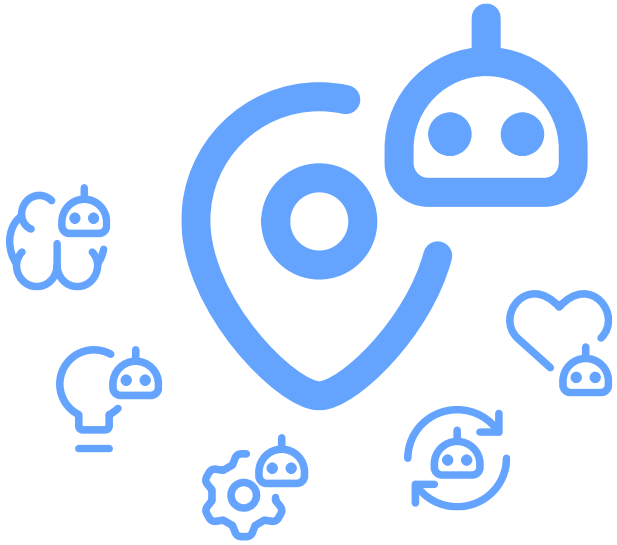
Maximilian Schreiner
KI Pro, The Decoder

Matthias Bastian
KI Pro, The Decoder

the decoder heise KI PRO heise I/O

Heise KI PRO ist ihr **Navigator** im KI-Alltag.

heise KI PRO



Die Lern- und Wissenscommunity für KI

Lernen. Vernetzen. Umsetzen.

heise KI PRO – Die Lern- und Wissenscommunity für KI

- ✓ **12 umfangreiche Fachartikel (Deep Dives)** zu aktuellen KI-Themen mit Fokus auf Einsatz im Unternehmen, inklusive Handlungsempfehlungen
- ✓ **12 Webinare** zu den jeweiligen Fachartikeln mit Diskussion
- ✓ **12 KI-PRO-Talks** zu aktuellen Themen und Interaktion mit der Community
- ✓ **24 Newsletter** (KI-Business-Briefings) mit weiterführenden Themen zu den Fachartikeln und kuratierter Leseliste
- ✓ **Zugang zur KI PRO Community Plattform** mit Forum, Fachartikel- und Newsletter-Archiv, Webinar-Aufzeichnungen, direkten Kontakt zur Redaktion, Einsendemöglichkeiten für Fragen, Erfahrungsaustausch mit anderen Unternehmen
- ✓ **Zugang zum KI-Tool heise I/O sowie regelmäßige Consulting-Termine**
- ✓ **1x heise+/Pur** mit Zugang zu allen Inhalten aus heise online, c't, iX, Mac & i, Make und c't Fotografie

heise KI PRO – Deep Dive Bibliothek

- ✓ Prompt Engineering 34 Seiten
- ✓ Generative KI-Tools 92 Seiten, 36 der besten Tools
- ✓ KI und Urheberrecht 19 Seiten
- ✓ KI-Richtlinien 19 Seiten + Checkliste & Muster
- ✓ KI-Governance 23 Seiten
- ✓ EU AI Act 25 Seiten
- ✓ KI-Change-Management 27 Seiten
- ✓ Cases & Ergebnisse aus dem Alltag der heise Gruppe 25 Seiten
- ✓ KI-Barometer Januar 2025 26 Seiten
- ✓ KI-Agenten 47 Seiten
- ✓ KI-Modellkunde 65 Seiten
- ✓ KI-Modellkunde: Bildmodelle 34 Seiten
- ✓ KI-Workflows 31 Seiten
- ✓ Deep Research 36 Seiten

Agenda

heise KI PRO

Webinar: KI-Halluzinationen

Was sind KI-Halluzinationen?

Wie entstehen KI-Halluzinationen?

Welche Arten von KI-Halluzinationen gibt es?

Wie erkenne ich KI-Halluzinationen?

Welche Strategien gibt es gegen KI-Halluzinationen?

heise KI PRO

Was sind KI-Halluzinationen?

Definition

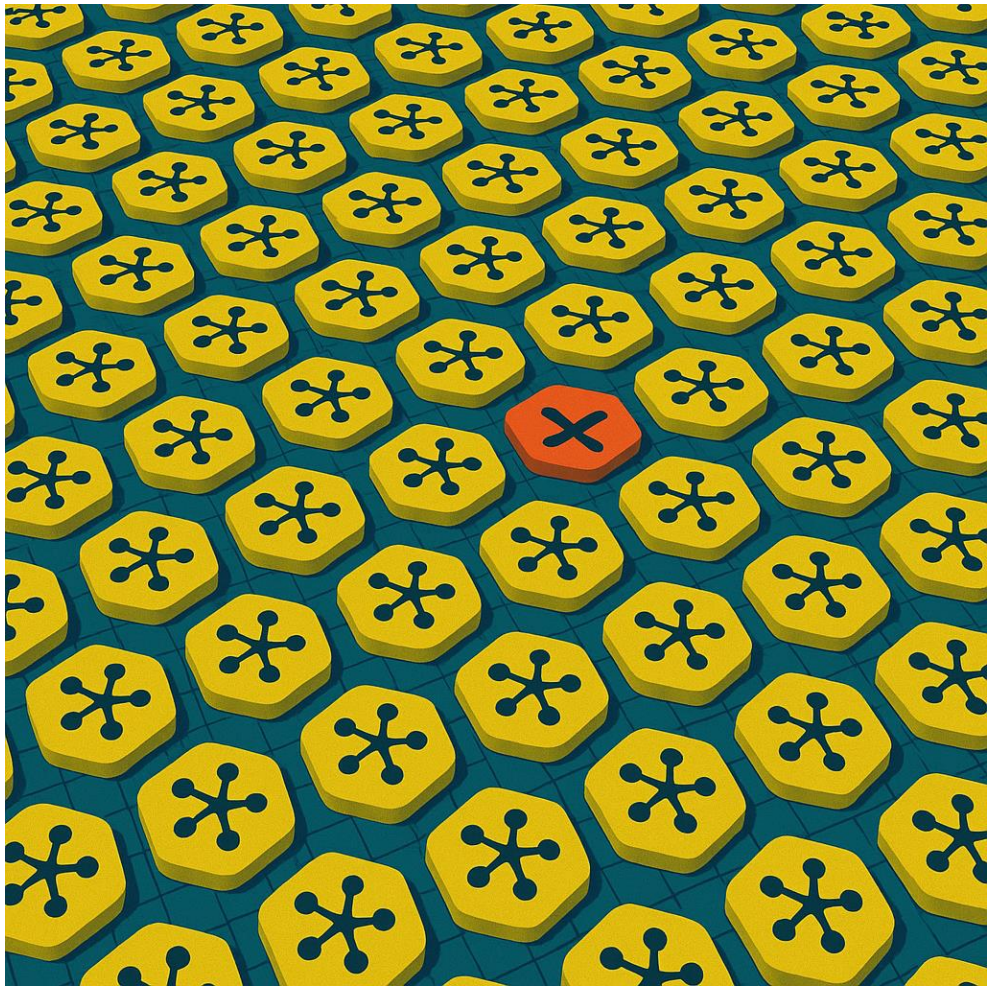
heise KI PRO

KI-Halluzinationen sind falsche oder irreführende Ergebnisse, die von KI-Modellen generiert werden.

Deloitte Australia

heise KI PRO

Deloitte Australia



Fall: Deloitte Australia lieferte einen Prüfbericht mit KI-generierten, falschen Fußnoten an die australische Regierung.

Finanzielle Folge: Rückerstattung von Teilen von 439.000 AUD (ca. 245.000 €).

Hauptschaden: Immenser Verlust an Reputation und Vertrauen.

KI-Halluzinationen haben reale geschäftliche Konsequenzen. Qualitätssicherung und menschliche Kontrolle sind unerlässlich.

Bilder und Halluzinationen

heise KI PRO

Bilder und Halluzinationen

heise KI PRO

heise KI PRO

Wie entstehen KI-Halluzinationen?

heise KI PRO

Demo

Wie entstehen KI-Halluzinationen?

heise KI PRO

		01.02.1967	0,00003847291 %
		02.02.1967	0,00007652143 %
		03.02.1967	0,00001298567 %
		04.02.1967	0,00005467892 %
An welchem Tag hat Frau Wilma Maier Geburtstag?	Frau Wilma Maier hat am	05.02.1967	0,00009834721 %
		06.02.1967	0,00006359274 %
		07.02.1967	0,00004625187 %
		08.02.1967	0,00008741329 %
		09.02.1967	0,00002193856 %

heise KI PRO

Welche Arten von KI-Halluzinationen gibt es?

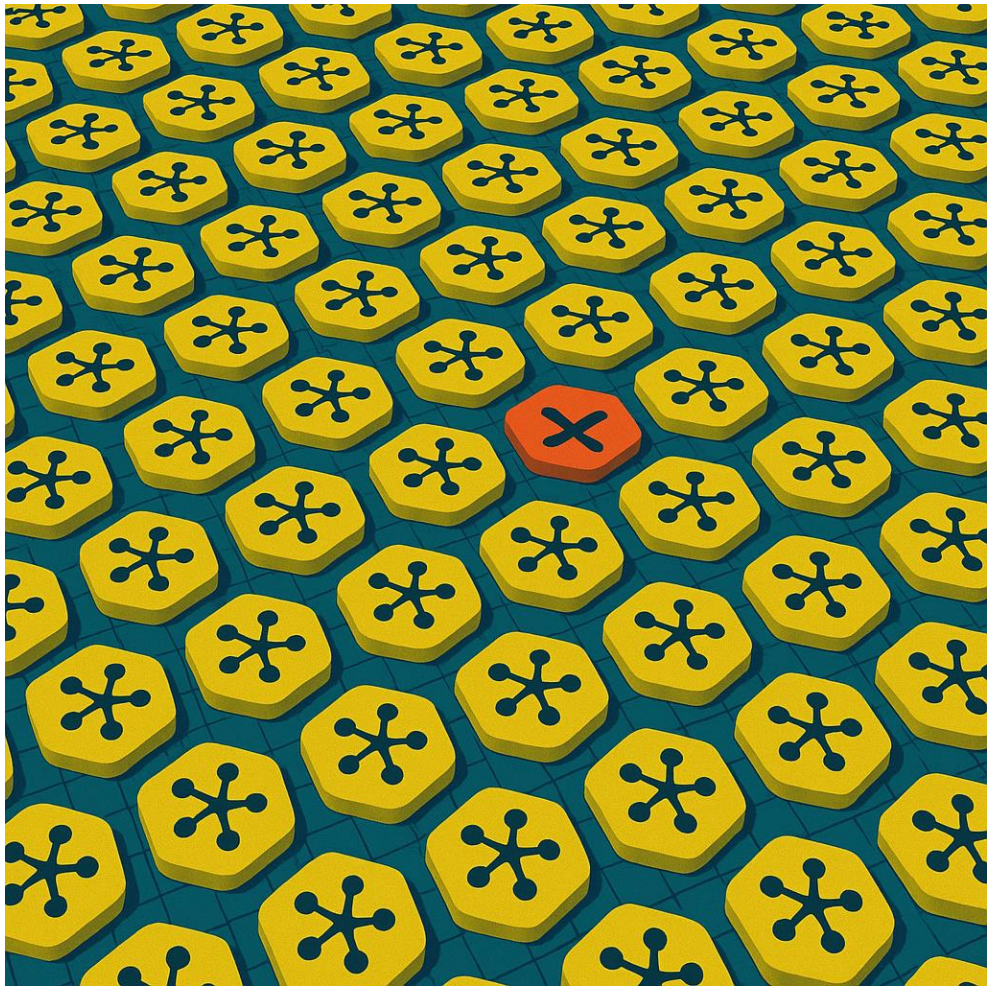
Welche Arten von KI-Halluzinationen gibt es?

heise KI PRO

Art der Halluzination	Beschreibung
Faktisch	Die Ausgaben sind falsch oder erfunden.
Zeitlich	Veraltetes oder überholtes Wissen wird als aktuell dargestellt.
Kontextuell	Fügt Konzepte hinzu, die im ursprünglichen Kontext nicht erwähnt oder impliziert wurden.
Linguistisch	Grammatikalisch korrekte, aber inhaltlich unsinnige oder unzusammenhängende Sätze.
Extrinsisch	Aussagen, die nicht durch die zugrunde liegenden Quellen gestützt sind.
Intrinsisch	Widersprüchliche oder sich selbst widersprechende Antworten.

Wie erkenne ich KI-Halluzinationen?

heise KI PRO

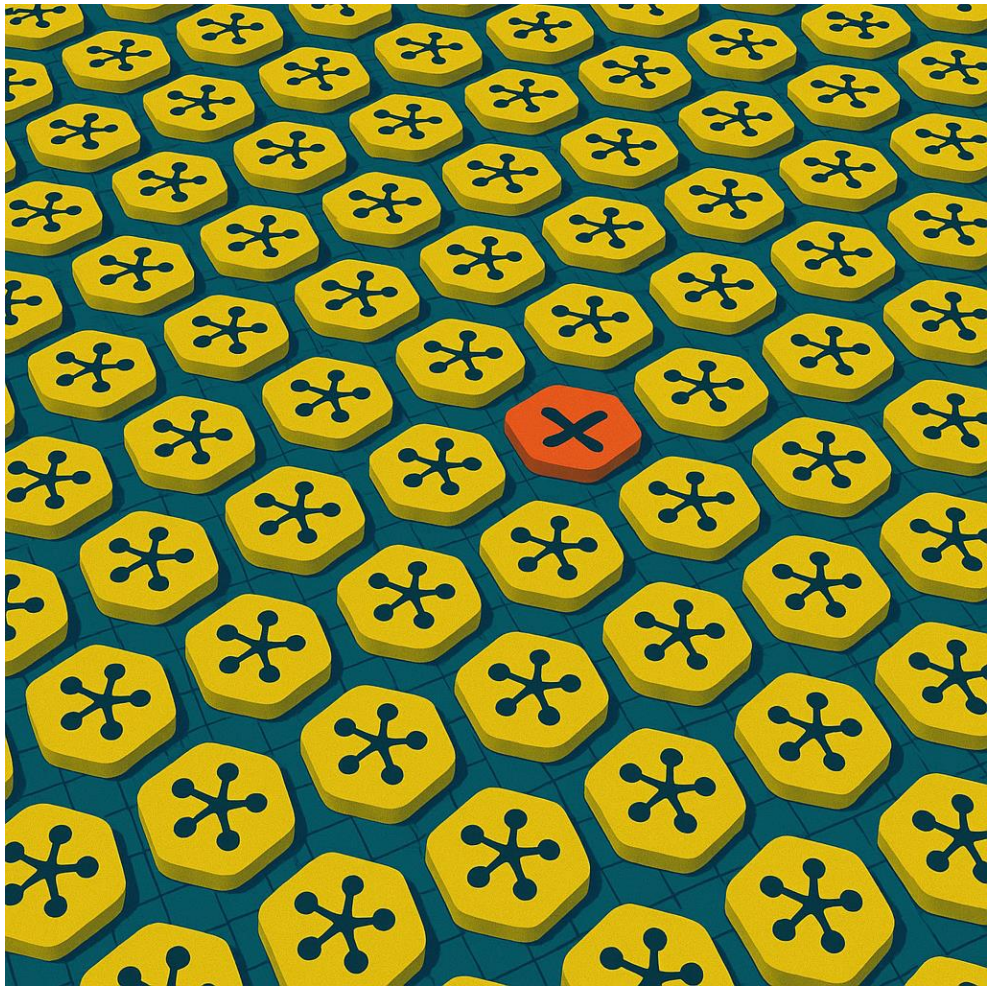


Warnsignale:

- Übertrieben selbstsicher bei schwierigen Themen
- Keine konkreten Quellen
- Verschiedene Antworten bei gleicher Frage
- Zu perfekte Details
- Zu einfache Lösungen für komplexe Probleme
- Detaillierte Infos zu obskuren Themen
- ...

Wie erkenne ich KI-Halluzinationen?

heise KI PRO

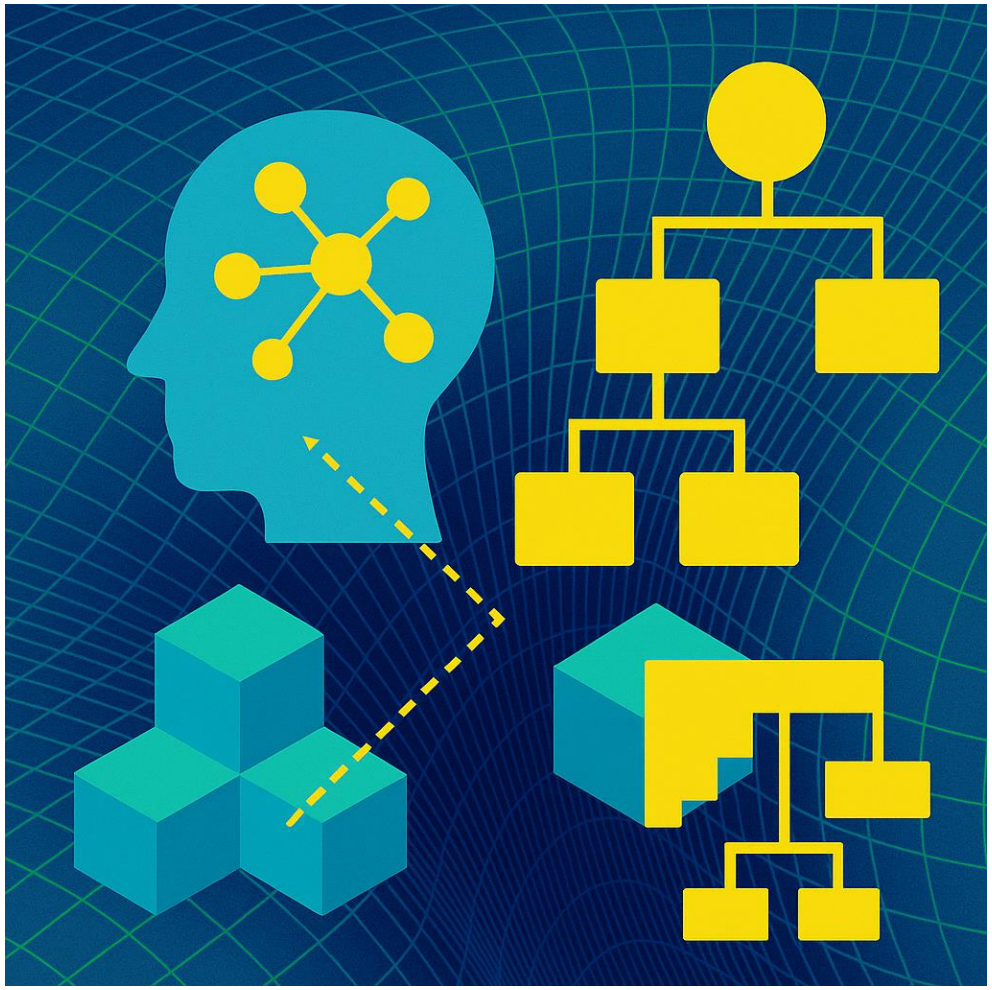


Damit man Halluzinationen verlässlich bekämpft und erkennt (und nicht nur „Warnsignale“ abliest), braucht es ein paar Voraussetzungen – auf Personen-, Prozess- und Systemebene.

- Domänenwissen / Urteilsfähigkeit
- Quellenkompetenz
- Fehlerkultur
- …

heise KI PRO

Welche Strategien gibt es gegen KI-Halluzinationen?



Die wichtigste Strategie:

Gut ausgebildete Menschen, die ihre Themen (Kontext) kennen, ihre Prompts und die Stärken und Schwächen von KI.

Kontext-Arbeit – Kontrolle über Wissen und Quellen

Halluzinationen entstehen häufig, wenn Modelle auf unklare, veraltete oder unvollständige Informationen zurückgreifen.

Context-Arbeit bedeutet, den Wissensraum aktiv zu gestalten – also sicherzustellen, dass die KI nur auf geprüfte, relevante und nachvollziehbare Informationen zugreifen kann.

Ziele:

- Kontrolle über die verwendeten Datenquellen behalten
- Transparenz und Nachvollziehbarkeit schaffen
- Den Einfluss des unsicheren Trainingswissens minimieren

Kontext-Arbeit – Kontrolle über Wissen und Quellen

heise KI PRO

Zentrale Methoden & Strategien:

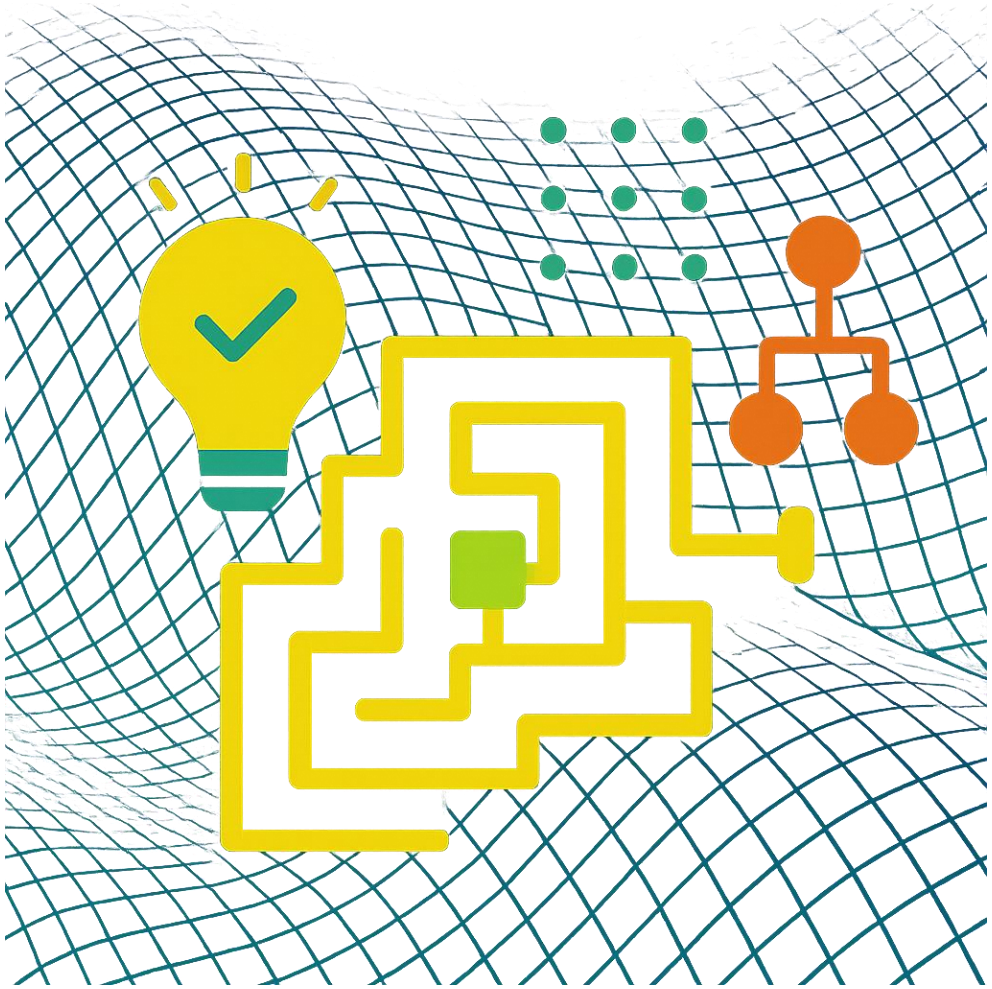
- **Context Engineering:** Relevante Dokumente, Daten oder Wissensquellen direkt im Chat bereitstellen
- **RAG-Systeme & Tool Use:** Externe Datenquellen oder APIs für aktuelle Informationen einbinden
- **Grounding über Zitate:** KI-Antworten müssen sich auf wörtliche Textstellen beziehen
- **Fact-Checking (LLM-as-a-Judge):** Ergebnisse durch Modelle oder Webquellen gegenprüfen
- **Fachgrenzen respektieren:** Expert:innen prüfen, was nur Fachwissen bewerten kann

Leitprinzip:

„Ich kenne meine Quellen – und die KI arbeitet nur innerhalb dieses Rahmens."

Prompting – Denken in klaren Aufgabenstrukturen

heise KI PRO



Die Qualität einer KI-Antwort hängt maßgeblich von der Klarheit und Struktur des Prompts ab.

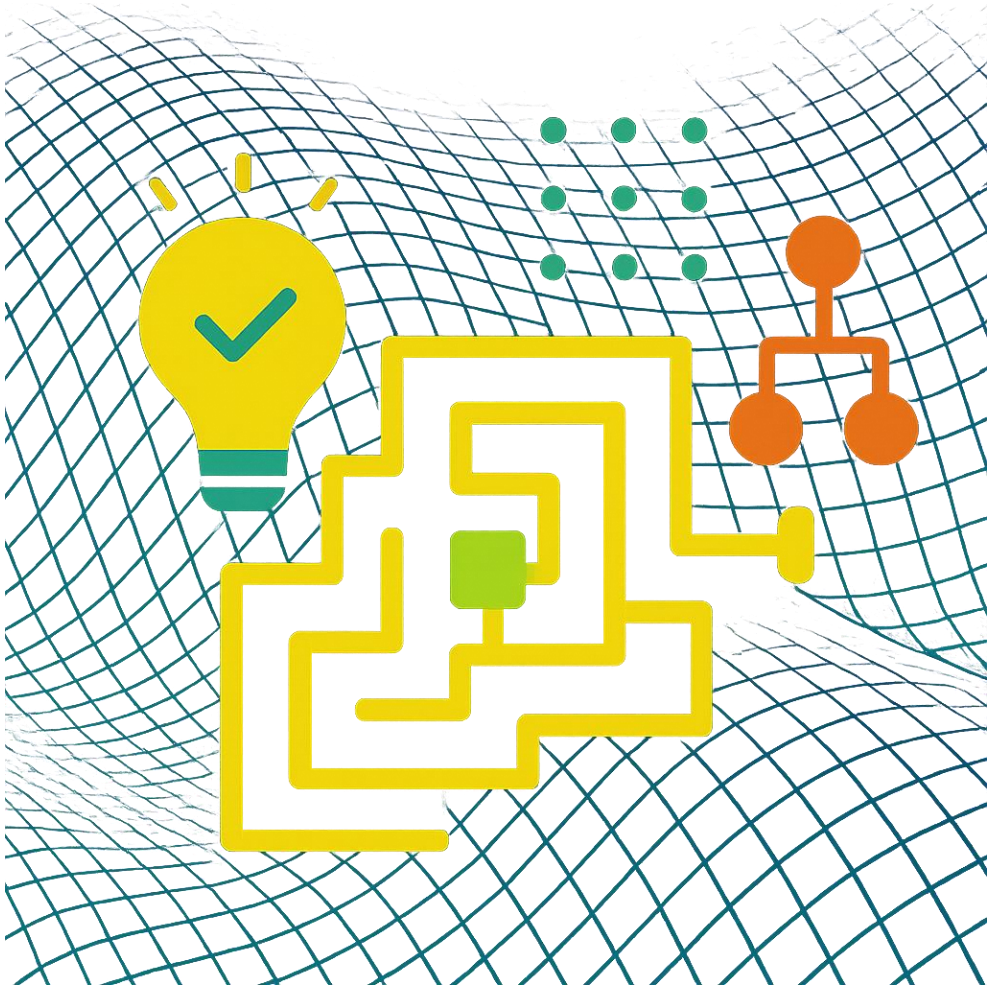
Gezieltes Prompting heißt, Aufgaben präzise zu formulieren, Erwartungen zu kommunizieren und das Modell zur Reflexion anzuleiten.

Ziele:

- Missverständnisse vermeiden
- Plausible, nachvollziehbare Antworten fördern
- Den Denkprozess der KI transparent machen

Prompting – Denken in klaren Aufgabenstrukturen

heise KI PRO



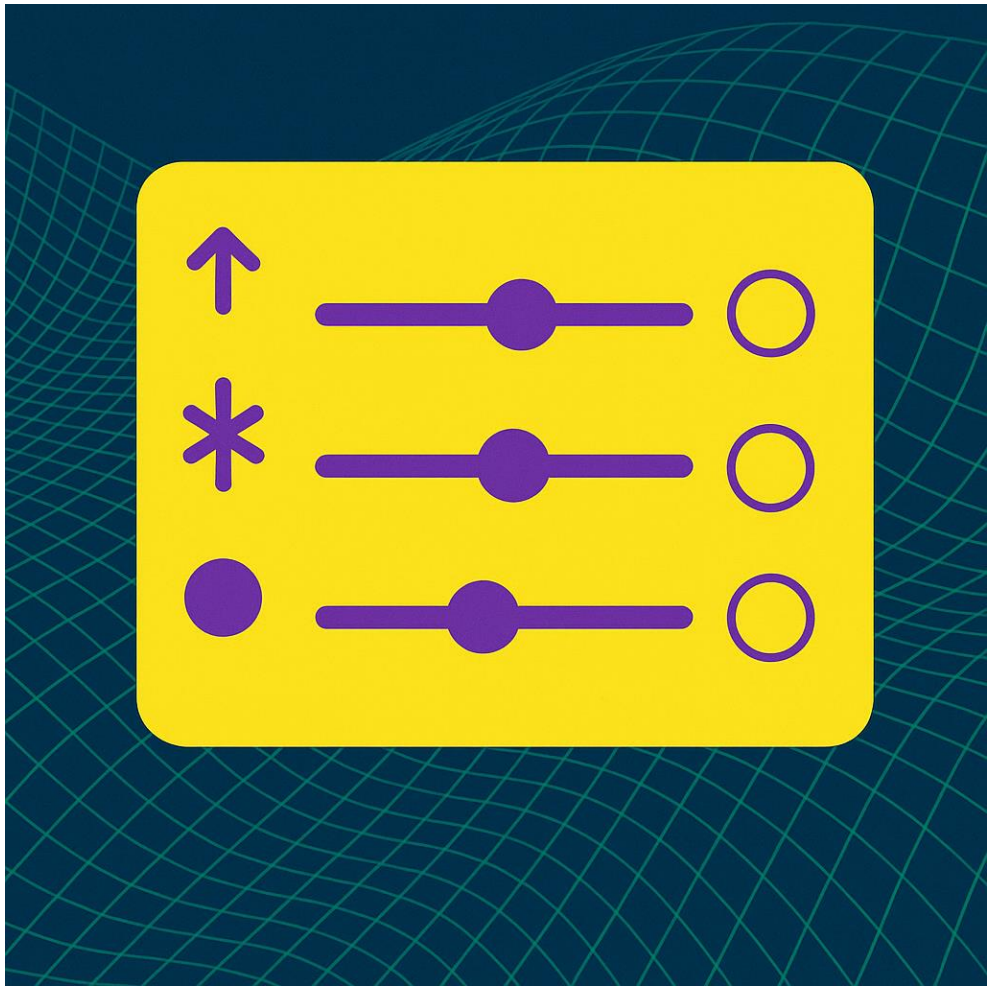
Zentrale Methoden & Strategien:

- **Chain-of-Thought:** KI legt ihre Denkschritte offen, bevor sie antwortet
- **„Ich weiß es nicht“-Erlaubnis:** Unsicherheit ist erlaubt und wird explizit gefördert
- **Multi-Shot Prompting:** Gute und schlechte Antwortbeispiele vorgeben
- **Konfidenz-Level einfordern:** Aussagen nach Sicherheitsgrad strukturieren
- **Praxiswissen nutzen:** Eigene Erfahrung mit Prompts und Modellverhalten einfließen lassen

Leitprinzip:

„Ich gestalte meine Prompts so, dass die KI nicht rät – sondern denkt, prüft und belegt.“

Technische Lösungen – Systeme stabil und nachvollziehbar halten



Neben Kontext und Kommunikation spielen auch technische Parameter und Kontrollmechanismen eine entscheidende Rolle.

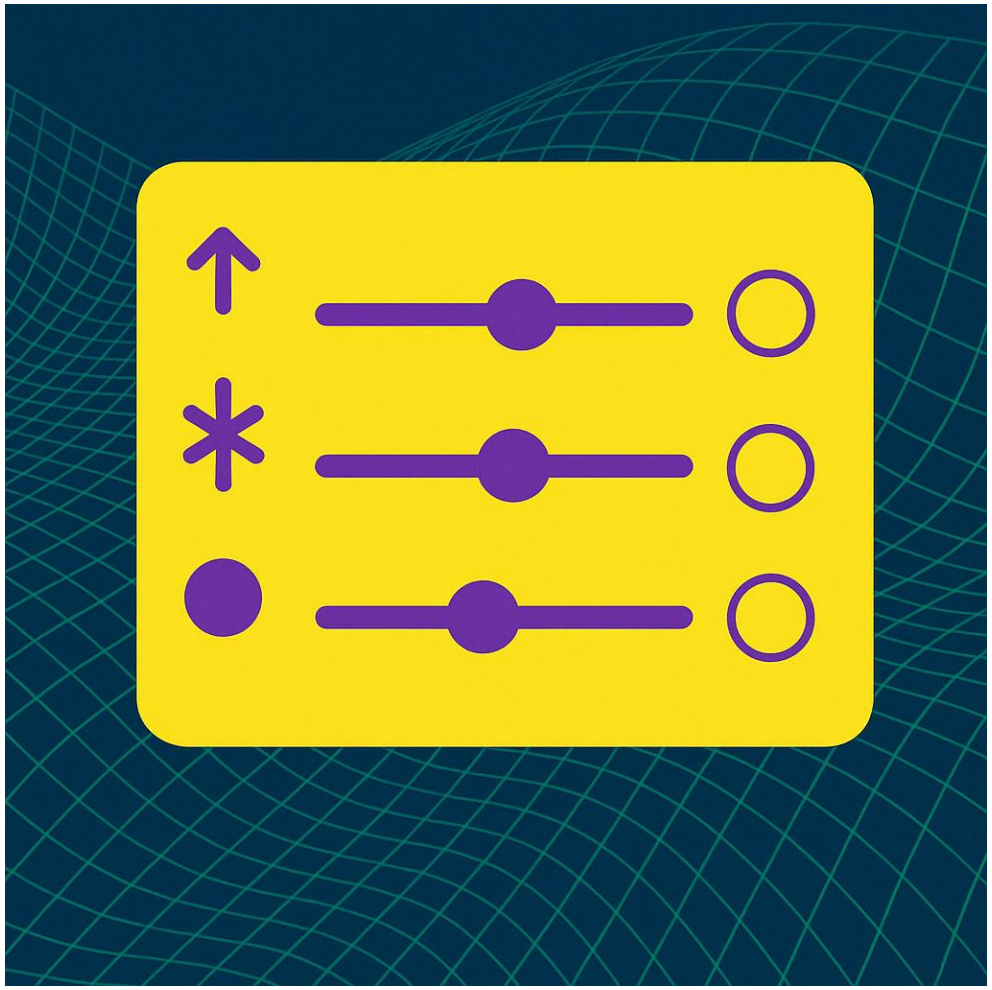
Technische Lösungen können helfen, Fehlerquellen systematisch zu erkennen und zu begrenzen – besonders bei großflächigem oder automatisiertem KI-Einsatz.

Ziele:

- Stabilität und Konsistenz der Ausgaben erhöhen
- Fehlverhalten frühzeitig erkennen
- Automatisierte Prüfmechanismen implementieren

Technische Lösungen – Systeme stabil und nachvollziehbar halten

heise KI PRO



Zentrale Methoden & Strategien:

- **Best-of-N-Verifikation:** Mehrere Modellläufe vergleichen, Inkonsistenzen aufdecken
- **Guardrails:** Technische Schutzschichten (z. B. Filter, Quellvalidierung, Schwellenwerte) einsetzen
- **Temperatur-Kontrolle:** Kreativität und Variation gezielt steuern (niedrig für Fakten, hoch für Ideen)
- **Monitoring & Evaluation:** Ergebnisse regelmäßig prüfen und auswerten

Leitprinzip:

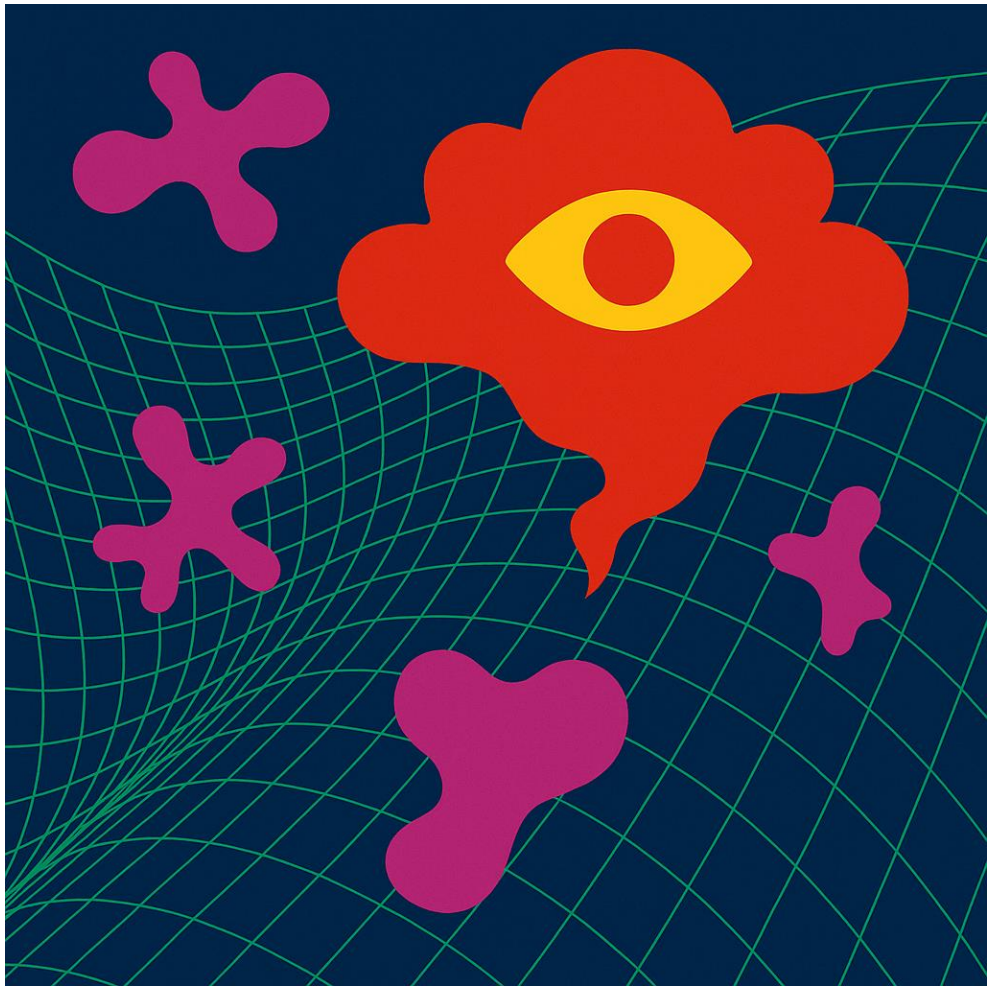
„Ich nutze technische Kontrollen, um Verlässlichkeit und Transparenz zu sichern.“

heise KI PRO

Zusätzliche Risikofaktoren

Es gibt einige spezielle Risikoquellen, die Sie kennen sollten.

heise KI PRO



Spezielle Risikofaktoren für KI-Halluzinationen:

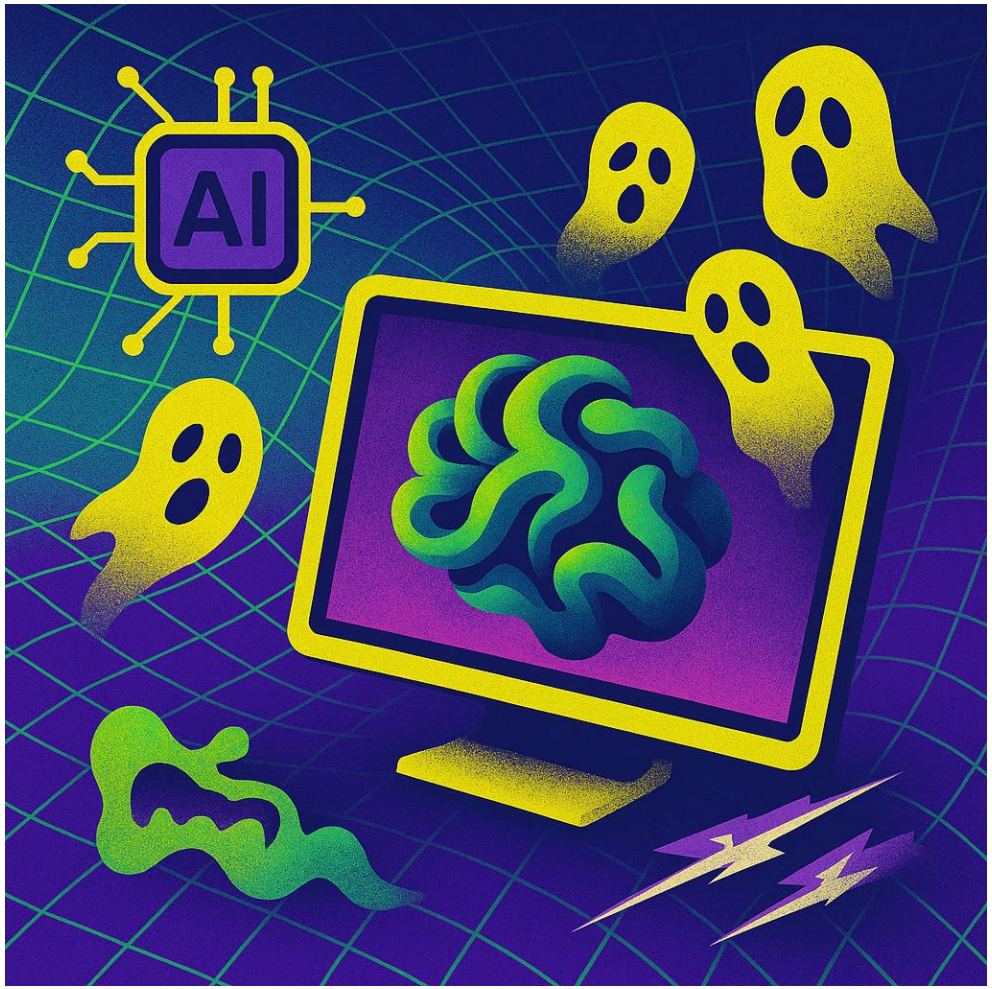
- Schmeichelei (Sycophancy)
- „Lost in the Middle“-Problem
- Token-basierte Rechenfehler
- „Erinnerungs“-Funktionen

heise KI PRO

Fazit

Mit KI arbeiten, heißt mit Risiko arbeiten.

heise KI PRO



Keine Angst for Fehlern!

Wie wichtig ist Akkuratheit?

Wie schwierig sind Fehler zu finden?

Benutzen Experten und Expertinnen den KI-Prozess?

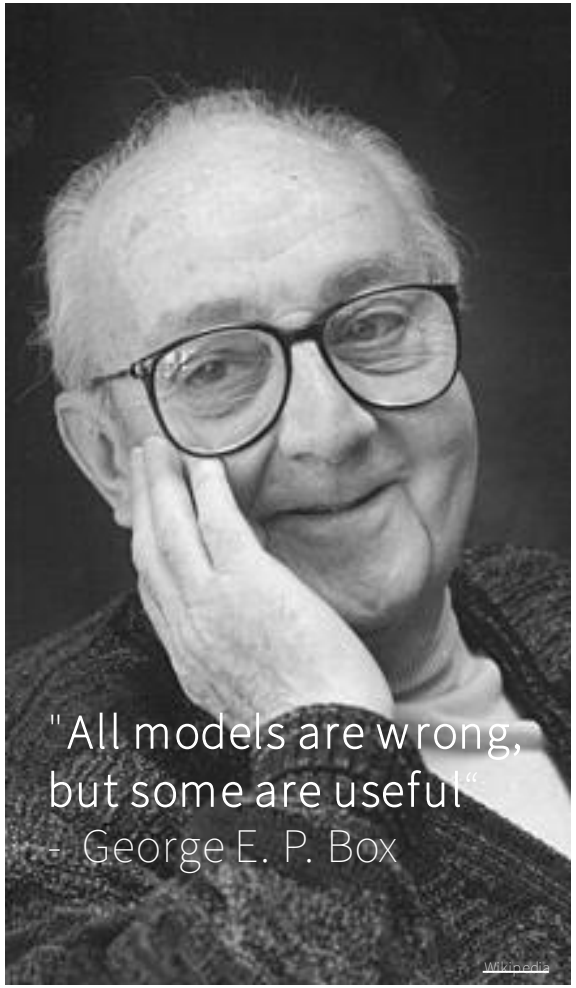
Könnte ich bei Menschen höhere Akkuratheit erwarten?

heise KI PRO

Das größte Risiko generativer KI ist die selbstgewählte Unmündigkeit:

- Wenn wir der KI blind vertrauen, statt sie zu führen.
- Wenn wir verlernen, kritisch zu denken und zu hinterfragen.
- Wenn wir Bequemlichkeit über Urteilskraft stellen.

Ihre Aufgabe:
Werden Sie zum kritischen Partner Ihrer KI.



heise KI PRO

Fragen?

the decoder heise KI PRO heise I/O

heise KI PRO – Termine

⊠ KI PRO Webinare

- 21. Oktober 2025 (16:00 - 17:00 Uhr)
  KI-Modellkunde: Fundierte Entscheidungshilfe für KI-Modelle im Unternehmen [WDH]
- 18. November 2025 (16:00 - 17:00 Uhr)
  KI-Guidelines im Unternehmen - (Rechtlich) notwendig oder nicht? [WDH]
- 11. November 2025 (16:00 - 17:00 Uhr)
  ⊠ KI im Büro: Excel, Präsentationen und E-Mails mit KI erstellen
- 16. Dezember 2025. (16:00 - 17:00 Uhr)
  KI-Modellkunde: Fundierte Entscheidungshilfe für Bildmodelle im Unternehmen [WDH]

⊠KI PRO Talks

- 28. Oktober 2025 (16:00 - 17:00 Uhr)
  KI PRO TALK | Community Edition
- 25. November 2025
  KI PRO TALK | Community Edition
- 30. Dezember 2025
- KI PRO TALK | Community Edition

heise KI PRO

Benjamin Danneberg
Leitung | heise KI PRO
ben@deep-content.io

Kim M. Scheurenbrand
Manager | heise KI PRO
kim@deep-content.io

Unsere KI-Workflow Software für Teams
heise-io.de

Fach-Community
heise KI PRO
pro.heise.de/ki

the decoder heise KI PRO heise I/O