mr.petz AI
Make Business Intelligent

WAS ICH SCHON IMMER WISSEN WOLLTE · HEFT 2 · SEPTEMBER 2026 · KW 38/39

Woher weiß ich, ob ChatGPT mich gerade anlügt?

Antwort: am Tonfall nicht. Falsche Antworten klingen genauso sicher wie richtige — das ist gemessen. Auf den nächsten Seiten steht, woran Sie es trotzdem erkennen, welche Prüfung zwei Minuten dauert und welche Prüfung nichts bringt, obwohl sie fast alle machen.

Aus der Praxis für die Praxis.

Zahlen
Wie oft es danebengeht

Quellen
Erfunden oder echt

Prüfen
Zwei Minuten, drei Griffe

Ernstfall
Was Gerichte dazu sagen

Christian Niebisch · Head of AI
mrpetzai.de · Stand 4. September 2026

VIER UNTERSUCHUNGEN

FÜNF DOKUMENTIERTE FÄLLE

2

Wie oft liegt eine KI eigentlich daneben?

Öfter, als der Ruf vermuten lässt — und deutlich öfter, als die Hersteller selbst messen. Vier unabhängige Untersuchungen aus den letzten anderthalb Jahren, jede mit anderer Frage, jede mit anderem Ergebnis. Was sie gemeinsam haben, steht unter der Tabelle.

WER HAT GEMESSEN	WAS GENAU	ERGEBNIS
Europäische Rundfunkunion 22 Sender, 18 Länder, 14 Sprachen, Oktober 2025. ARD, ZDF und Deutsche Welle waren dabei	Über 3.000 Antworten von ChatGPT, Copilot, Gemini und Perplexity zu Nachrichtenfragen, bewertet von Journalisten	**45 % der Antworten hatten mindestens einen erheblichen Mangel.** Bei Gemini waren es 76 %
Columbia Journalism Review März 2025	1.600 Abfragen: Ein echter Textauszug wurde vorgelegt, gefragt war nach Überschrift, Verlag und Adresse	**Über 60 % falsch.** Bei ChatGPT 134 von 200 Antworten
Stiftung Warentest Januar 2026	40 Alltagsfragen aus Geld, Recht, Gesundheit und Verbraucherrechten	Richtig beantwortet: Perplexity 72 %, Gemini 68 %, Copilot 67 %, ChatGPT 65 %
Vectara-Rangliste Mai 2026	Die leichteste denkbare Aufgabe: einen mitgelieferten Text zusammenfassen, ohne eigenes Wissen	Selbst hier 2 bis 5 % Widersprüche zum Ausgangstext. **Nicht null**

DIE EINE ZAHL, DIE MAN SICH MERKEN SOLLTE

Bei Alltagsfragen zu Geld und Recht ist etwa jede dritte Antwort falsch — und zwar beim Marktführer. Nicht bei einem Nischenanbieter, nicht bei einer alten Version.

Und die Richtung stimmt nicht: NewsGuard misst seit 2024, wie oft Chatbots nachweislich falsche Behauptungen wiederholen. August 2024: 18 %. August 2025: 35 %. Der Grund ist unangenehm — die Systeme verweigern die Antwort heute praktisch nie mehr. Sie sagen lieber irgendetwas.

Was sagen OpenAI, Anthropic, Google und Microsoft eigentlich selbst dazu?

Erstaunlich offen — nur steht es an einer Stelle, die niemand liest. Alle vier schreiben dasselbe in ihre Nutzungsbedingungen, und es ist kein Hinweis, sondern eine Pflicht.

ANBIETER	WORTLAUT, SINNGEMÄSS ÜBERSETZT
OpenAI ChatGPT, EU-Bedingungen, Stand 16.01.2026	„Sie müssen die Ausgabe auf Richtigkeit prüfen, einschließlich einer menschlichen Durchsicht, bevor Sie sie verwenden oder weitergeben." Und: nicht als alleinige Wahrheitsquelle verwenden.
Anthropic Claude, Hilfebereich, Stand 16.03.2026	Erfundene Zitate „mögen maßgeblich aussehen oder überzeugend klingen, sind aber nicht in Tatsachen begründet". Nutzer sollen Claude „nicht als einzige Wahrheitsquelle" behandeln.
Google Gemini, deutsche Hilfeseite	Im Original auf Deutsch: „Gemini kann KI-Halluzinationen ausgeben und unrichtige Informationen als Fakten darstellen." Und: „Überprüfen Sie daher die Antworten."
Microsoft Copilot, Nutzungsbedingungen, Stand 18.08.2026	„Manchmal sind die Quellen, die Copilot verwendet, nicht verlässlich, nicht einschlägig oder nicht richtig." Und: „Wir sind nicht verantwortlich für Folgen, die aus Ihrem Vertrauen auf Copilot entstehen."

VERLASSEN SIE SICH NICHT AUF DEN WARNHINWEIS IN DER OBERFLÄCHE

Microsoft hat den Hinweis „KI-generierte Inhalte können falsch sein" in Microsoft 365 Copilot Ende 2025 standardmäßig ausgeblendet. Begründung: Manche Nutzer fanden ihn zu störend, andere zu unauffällig. Ein fehlender Warnhinweis heißt also nicht, dass geprüft wurde.

Das Landgericht München I hat im Mai 2026 außerdem entschieden, dass ein allgemeiner KI-Warnhinweis den Anbieter nicht vor Haftung schützt. Für Sie heißt das umgekehrt: Er schützt Sie auch nicht.

4

Woran erkenne ich eine erfundene Quelle?

Daran, dass Sie sie anklicken. Das klingt banal, ist aber der ganze Trick — denn eine erfundene Quelle sieht im Text vollkommen echt aus. Sie hat ein Datum, ein Aktenzeichen, einen Verlag und oft sogar einen realen Autorennamen.

Die drei Fehlerbilder, gemessen an 1.600 Abfragen

1 Die Adresse existiert nicht

Der Link führt auf eine Fehlerseite. Bei über der Hälfte der Antworten von Gemini und Grok waren die angegebenen Adressen erfunden oder tot.

2 Die Quelle trägt die Aussage nicht

Der häufigere und gefährlichere Fall: Der Link funktioniert, die Seite ist seriös — nur steht dort etwas anderes, als behauptet wurde.

3 Kein Zögern, nirgends

ChatGPT zeigte bei 134 falschen Antworten nur 15-mal Unsicherheit an. Keines der acht getesteten Werkzeuge hat je eine Antwort verweigert.

Vier Griffe, die eine Fälschung auffliegen lassen

- Den Link wirklich anklicken. Nicht überfliegen — öffnen. Nicht geöffnet heißt nicht geprüft.
- Prüfen, ob dort steht, was behauptet wurde — nicht nur, ob die Seite lädt.
- Gegen ein Verzeichnis prüfen, nicht gegen die KI. Paragrafen auf gesetze-im-internet.de, Tarifverträge beim Verband, Fristen in der Originalvorschrift.
- Namen sind kein Beleg. In einem dokumentierten Fall waren die zitierten Wissenschaftler echt — die Zitate waren erfunden.

5

Vier Griffe, die eine Fälschung auffliegen lassen

DER SATZ, DER AUF DIESER SEITE ZÄHLT

Der Tonfall einer Antwort sagt nichts über ihre Richtigkeit. Eine erfundene Auskunft klingt genauso ruhig und sachlich wie eine geprüfte. Sie können sich Ihr Bauchgefühl an dieser Stelle sparen — es hilft hier nicht.

KAPITEL 4 · DER HÄUFIGSTE KUNSTFEHLER

Warum bringt „Bist du sicher?“ überhaupt nichts?

Weil das Programm dann nicht nachprüft, sondern nachgibt. Das ist untersucht worden, und das Ergebnis ist deutlich genug, um die Gewohnheit abzulegen.

WAS PASSIERT, WENN SIE WIDERSPRECHEN

Acht verbreitete Systeme wurden geprüft. Sobald die Nutzerin widersprach, wechselten sie in 56 % der Fälle auf die vorgehaltene Gegenposition. Entscheidend ist der zweite Wert: Lag das Modell ursprünglich richtig, gab es seine richtige Antwort in 45 % der Fälle auf und ersetzte sie durch eine falsche.

Am wirksamsten war dabei nicht das gute Argument, sondern der schlichte, selbstbewusste Einwand ohne Begründung — der überzeugte in 85 % der Fälle.

UND DIE RÜCKFRAGE NACH DER EIGENEN QUELLE?

Genauso wertlos. In dem Verfahren, mit dem das Thema 2023 weltweit bekannt wurde, fragten die Anwälte ChatGPT ausdrücklich, ob die zitierten Urteile echt seien. ChatGPT versicherte, sie existierten und seien in den einschlägigen Datenbanken zu finden. Beides war falsch.

Was stattdessen funktioniert

- Neuer, leerer Chat. Dieselbe Frage noch einmal stellen, ohne die erste Antwort zu erwähnen. Weichen die beiden Antworten voneinander ab, ist mindestens eine falsch.
- Beide Antworten nebeneinanderlegen und neutral bewerten lassen, statt eine davon anzugreifen. In der Untersuchung war das deutlich stabiler als der Widerspruch im laufenden Gespräch.
- Ein zweites Werkzeug fragen. Die Fehler sind herstellerabhängig: Wo Gemini in der Rundfunkstudie bei 76 % Beanstandungen lag, lag Perplexity bei 30 %. Zwei verschiedene Anbieter irren selten gleich.

KAPITEL 5 · ZUM KOPIEREN

Wie prüfe ich das in zwei Minuten?

Der graue Text ist zum Abtippen oder Kopieren. Alles in eckigen Klammern ersetzen Sie durch Ihres. Alle vier laufen in der kostenlosen Version.

Neuer, leerer Chat, dieselbe Frage: Weichen die Antworten ab, ist mindestens eine falsch.

Zwei weitere Griffe zum Kopieren

1 VOR DER FRAGE

Unwissen ausdrücklich erlauben

> „Beantworte die folgende Frage. Wenn du dir bei einem Punkt nicht sicher bist oder die Angaben fehlen, schreibe wörtlich: Dazu habe ich keine ausreichenden Informationen. Rate nicht. Frage: [Ihre Frage]“

Warum das wirkt: OpenAI schreibt selbst, die üblichen Trainingsverfahren „belohnen Raten und bestrafen das Eingeständnis von Unsicherheit“. Die Erlaubnis, nichts zu wissen, müssen Sie also ausdrücklich erteilen — von allein nimmt sie sich das Programm nicht.

2 BEI EIGENEN UNTERLAGEN

Auf den mitgelieferten Text festnageln

> „Antworte ausschließlich auf Grundlage des folgenden Textes. Zitiere zu jeder Aussage die Stelle wörtlich, auf die du dich stützt. Findest du für eine Aussage keine Stelle, lasse sie weg und schreibe dort [ohne Beleg]. Text: [einfügen]“

Was das bringt: Genau diese Aufgabe ist die leichteste, die es gibt — und die Fehlerquote sinkt dabei auf 2 bis 5 %. Nicht auf null. Die wörtlichen Zitate sind der Punkt: Sie können sie im Ausgangstext wiederfinden.

3 BEI BEHAUPTUNGEN MIT QUELLE

Die Quellenprüfung erzwingen

> „Nenne zu jeder Angabe die Quelle mit vollständiger Adresse und Datum. Gib nur Quellen an, die du tatsächlich aufgerufen hast. Wenn du eine Angabe nicht belegen kannst, kennzeichne sie als unbelegt.“

Danach: jede Adresse anklicken. Der Zusatz reduziert Erfindungen, er verhindert sie nicht.

4 DIE GEGENPROBE

Dieselbe Frage im neuen Chat

> „[Ihre Frage, wortgleich wie beim ersten Mal]“ — in einem neuen, leeren Chat, ohne Hinweis auf die erste Antwort.

Auswertung: Gleiche Antwort — vermutlich tragfähig. Andere Antwort — mindestens eine der beiden ist falsch, und Sie wissen nicht welche. Dann prüfen Sie von Hand.

Vier Griffe zum Abtippen: Unwissen erlauben, festnageln, Quelle erzwingen, gegenprüfen.

8

Warum rechnet ChatGPT falsch, obwohl es doch ein Computer ist?

Weil es nicht rechnet. Ein Sprachmodell sagt die wahrscheinlichste nächste Zeichenfolge voraus — auch bei Ziffern. Das ist kein Taschenrechner, der sich vertut, sondern ein Textprogramm, das eine Zahl für plausibel hält.

WIE DEUTLICH DAS IST

In einer Untersuchung von 2025 sollte GPT-4o zwei fünfstellige Zahlen multiplizieren — 30-mal. In 2 von 30 Fällen kam das richtige Ergebnis heraus. Die Fehler waren dabei keine Verständnisfehler, sondern stille Zahlendreher mitten in einer ansonsten plausiblen Herleitung. Teilweise war das Endergebnis falsch, obwohl jeder einzelne gezeigte Rechenschritt stimmte.

Woran Sie sehen, ob wirklich gerechnet wurde

WAS SIE SEHEN	WAS DAS HEISST	WAS SIE TUN
Die Zahl steht einfach im Fließtext	Sie wurde vorhergesagt, nicht gerechnet	Nachrechnen. Bei jeder Summe, die zählt
Es erscheint ein Rechen- oder Analysefeld, oder eine Datei wird erzeugt	Es lief ein echtes Rechenwerkzeug	Trotzdem eine Summe stichprobenartig prüfen — OpenAI verlangt das selbst
Sie haben die Tabelle als Bildschirmfoto eingefügt	Der schlechteste Weg von allen	Als Datei hochladen oder Werte eintippen

FÜR DAS PERSONALBÜRO BESONDERS WICHTIG

OpenAI schreibt ausdrücklich, dass exakte Werte aus Bildtabellen und eingescannten Dateien nicht zuverlässig ausgelesen werden. Gehaltslisten, Stundenzettel und Urlaubstabellen also nie abfotografieren und hineinziehen. Und selbst dann: Was gerechnet aussieht, ist erst geprüft, wenn Sie eine Zeile selbst nachgerechnet haben.

9

Was ist passiert, als jemand nicht nachgeprüft hat?

Fünf dokumentierte Fälle, vier davon aus dem Jahr 2026. Es traf keine überforderten Einzelpersonen, sondern Anwaltskanzleien, ein Gericht, einen Weltkonzern und eine der vier größten Prüfungsgesellschaften der Welt.

WANN UND WO	WAS PASSIERT IST	FOLGE
Kammergericht Berlin 20.11.2025	Im Schriftsatz einer Anwältin standen zwei Fundstellen — ein BGH-Beschluss und ein OLG-Beschluss — mit Datum und Aktenzeichen. Beide gibt es nicht.	Förmliche Ermahnung. Das Gericht spricht von einer „fantasierenden" KI
Amtsgericht Köln 02.07.2025	Ab Seite acht eines Schriftsatzes: erfundene Fundstellen, falsch zugeordnete Autoren, nicht existierende Aufsätze.	Verstoß gegen Berufsrecht; das Gericht nennt möglichen Prozessbetrug
Landgericht München I 28.05.2026	Googles KI-Übersicht behauptete über einen Münchner Verlag frei erfunden eine „Betrugsmasche" und unseriöse Geschäftspraktiken.	Google haftet dafür wie für eine eigene Aussage. Der KI-Warnhinweis half nicht
Air Canada Entscheidung 14.02.2024	Der Chatbot auf der eigenen Website nannte einem Kunden eine Tariffrist, die es so nicht gab. Die Airline argumentierte, der Chatbot sei eine eigenständige Einheit.	Zurückgewiesen: „Es macht keinen Unterschied, ob die Auskunft von einer Seite oder einem Chatbot kommt." Schadenersatz
Deloitte Australien Oktober 2025	Ein Gutachten für ein Ministerium, Auftragswert 440.000 AUD, enthielt über ein Dutzend erfundener Quellen — darunter erfundene Zitate real existierender Wissenschaftler.	Rückzahlung der Schlussrate; korrigierte Fassung mit Offenlegung des KI-Einsatzes

10

Drei weitere dokumentierte Fälle

Eine öffentliche Sammlung zählte am 3. September 2026 weltweit 2.009 Gerichtsverfahren, in denen ein Gericht förmlich festgestellt hat, dass sich Beteiligte auf erfundene KI-Inhalte gestützt haben. Elf davon aus Deutschland, zwei aus Österreich.

ZUM WEITERGEBEN

Wer haftet, wenn eine falsche KI-Auskunft nach draußen geht?

Diese Seite ist für die Geschäftsführung geschrieben. Drucken Sie sie aus und legen Sie sie hin — die Antwort auf die Überschrift ist nämlich: der Betrieb, nicht der Anbieter.

Drei Sätze, die die Rechtslage zusammenfassen

- **Die Prüfpflicht liegt vertraglich beim Nutzer.** Alle vier großen Anbieter schreiben in ihre Bedingungen, dass die Ausgabe vor Verwendung zu prüfen ist. OpenAI verlangt ausdrücklich eine menschliche Durchsicht.
- **Der Warnhinweis schützt nicht.** Das Landgericht München I hat einen allgemeinen KI-Warnhinweis im Mai 2026 nicht als Haftungsschutz anerkannt.
- **Eine KI-Auskunft im Kundenkontakt ist eine Aussage des Unternehmens.** So hat es das kanadische Tribunal im Fall Air Canada entschieden — der Chatbot ist Teil des Auftritts, nicht ein Dritter.

11

Was ein Betrieb daraus machen sollte — vier Punkte, keiner davon teuer

1 Schriftliche Regel

Wofür KI genutzt werden darf, wofür nicht, und wer gegenzeichnet, bevor etwas nach außen geht. Eine Seite genügt.

2 Vier-Augen-Prinzip nach außen

Alles, was den Betrieb verlässt — Kundenmail, Angebot, Website, Absage — liest ein Mensch gegen. Das ist keine Bürokratie, das ist die vertraglich geschuldete Prüfung.

3 Geschäftskonten statt Privatzugänge

Nur dort gibt es einen Auftragsverarbeitungsvertrag und kein Mitlernen. Bei kostenlosen Privatkonten bietet das keiner der vier Anbieter an.

4 Eine Stunde Einweisung

Die drei Prüfgriffe aus diesem Heft, einmal gemeinsam durchgegangen. Das ist der billigste Punkt auf dieser Liste und der mit der größten Wirkung.

Was den Betrieb verlässt, liest ein Mensch gegen — das ist keine Bürokratie, sondern die geschuldete Prüfung.

12

Woher stammen die Angaben in diesem Heft?

Alles am 4. September 2026 an der Originalquelle nachgelesen. Wo eine Zahl nicht zu belegen war, steht sie hier nicht — auch dann nicht, wenn sie gut geklungen hätte.

Woher die Angaben in diesem Heft stammen

ANBIETERBEDINGUNGEN
openai.com/policies/eu-terms-of-use · support.claude.com · support.google.com/gemini · microsoft.com/microsoft-copilot (Nutzungsbedingungen und Transparenzhinweis)

RUNDFUNKSTUDIE
Europäische Rundfunkunion, „News Integrity in AI Assistants", 22.10.2025 — ebu.ch

QUELLENANGABEN
Columbia Journalism Review / Tow Center, „AI Search Has a Citation Problem", 06.03.2025 — cjr.org

ALLTAGSFRAGEN
Stiftung Warentest, KI-Chatbots im Test, 16.01.2026 — test.de

ZUSAMMENFASSUNGEN
Vectara Hallucination Leaderboard, Stand 11.05.2026 — github.com/vectara

NACHGEBEN BEI WIDERSPRUCH
Kim u. a., „LLM Sycophancy Under User Rebuttal", arXiv 2509.16533

RECHENFEHLER
Zhang und Graf, „Mathematical Computation and Reasoning Errors by Large Language Models", 14.08.2025 — arXiv 2508.09932

GERICHTSFÄLLE
brak.de (KG Berlin, 20.11.2025) · lto.de (LG München I, 28.05.2026) · Moffatt v. Air Canada, 2024 BCCRT 149 · damiencharlotin.com/hallucinations

UND WENN SIE ES EINMAL GEMEINSAM DURCHGEHEN WOLLEN

Die vier Prüfgriffe von Seite 6 an Ihren eigenen Fällen — eine Stunde als Teams-Termin, ohne Kosten und ohne Verkaufsgespräch. Sie schicken mir drei Aufgaben aus Ihrem Alltag, wir gehen sie durch, Sie behalten die Vorlagen.

13

Wenn Sie es einmal gemeinsam durchgehen wollen

ANSPRECHPARTNER
Christian Niebisch · Head of AI

E-MAIL
christian.niebisch@mrpetzai.de

IM NETZ
www.mrpetzai.de

NÄCHSTES HEFT
28. September 2026

Dieses Heft ersetzt keine Rechtsberatung. Es fasst öffentlich zugängliche Angaben von Anbietern, Gerichten und Forschungseinrichtungen zusammen, Stand 4. September 2026.