STIVEN CATALYST

DRITTE SERIE · NOVEMBER 2026

# MANAGEMENT REVIEW

Management ohne Theater.




KI

## DIE QUELLEN DER KI PRÜFEN BEVOR ES RAUSGEHT

Sechs Gerichtsentscheidungen, die es nie gab, wie oft ein Chatbot seine Belege erfand, warum ein echtes Zitat trotzdem irreführen kann, was dafür gebaute Rechtswerkzeuge dennoch falsch machten, und eine Karte für die Prüfung vor dem Versand.

**18 %** der Werke, die GPT-4 in 42 kurzen Literaturübersichten zitierte, gab es nicht; bei GPT-3.5 waren es 55 % (getestet im April 2023).

WALTERS & WILDER, 2023

IN DIESER AUSGABE

TITELTHEMA

**Sechs Fälle, die es nie gab**

SEITE 03

WAS DIE FORSCHUNG SAGT

**Für das Recht gebaut, trotzdem falsch**

SEITE 06

WERKZEUG DER AUSGABE

**Die Karte vor dem Versand**

SEITE 08

NR. 96 · KI

# In dieser Ausgabe

Ein KI-Werkzeug schreibt in Sekunden einen flüssigen Text, mit Namen, Daten und Quellen, die richtig aussehen. Manche sind es nicht. Geht dieser Text unter dem Namen des Teams hinaus, haftet das Team dafür. Diese Ausgabe fragt, wie oft die Quellen falsch sind, auf welche Weise, und was eine Prüfung vor dem Versand abdecken muss.

Ein New Yorker Gericht sanktionierte Anwälte, deren Schriftsatz sechs nicht existierende Entscheidungen zitierte. In 84 Texten von 2023 waren 55 % der Belege von GPT-3.5 und 18 % der von GPT-4 erfunden, und viele echte hatten falsche Zahlen. Rechtswerkzeuge auf eigenen Datenbanken gaben auf 17 % bis 33 % schwieriger Fragen trotzdem falsche oder falsch belegte Antworten. Sechs Fragen an jede Quelle und eine Karte schließen die Ausgabe.

*Stiven Janaqi* HERAUSGEBER · STIVEN CATALYST


**03** TITELTHEMA
**Sechs Fälle, die es nie gab**

**04** DIE ZAHLEN
**636 Belege, einzeln geprüft**

**05** DAS MODELL
**Falsch oder falsch belegt**

**06** WAS DIE FORSCHUNG SAGT
**Für das Recht gebaut, trotzdem falsch**

**07** WIE ES GEMESSEN WIRD
**Jede Quelle öffnen**

**08** WERKZEUG DER AUSGABE
**Die Karte vor dem Versand**

**09** QUELLEN
**Quellen und Methode**


SO LESEN SIE DIESE AUSGABE

**Zahl**
Jede Zahl hat Quelle und Jahr unten auf ihrer Seite.

**Unsere Lesart**
Wo der Herausgeber deutet und nicht die Forschung, steht es dabei.

**Praxis**
Schritte und Karte sind Vorschläge zum Ausprobieren, keine Forschungsergebnisse.

TITELTHEMA

# Sechs Fälle, die es nie gab

Ein Passagier verklagte eine Fluggesellschaft wegen einer Knieverletzung durch einen Servierwagen. Gegen einen Antrag auf Klageabweisung zitierten seine Anwälte Gerichtsentscheidungen, die ChatGPT erzeugt hatte. Der unterzeichnende Anwalt hatte keine davon gelesen; er hatte den Text nur auf Stil geprüft.

### 1. März

Der Schriftsatz mit den Entscheidungen geht ans Gericht.

### 15. März

Die Anwälte der Fluggesellschaft schreiben, sie fänden die meisten Fälle nicht. Niemand zieht den Schriftsatz zurück.

### April

Das Gericht verlangt Kopien; die Anwälte reichen Auszüge ein und sagen, ein Fall sei nicht auffindbar.

### 25. Mai

Erstes Eingeständnis, dass ChatGPT die Fälle erzeugt hat.

### 22. Juni

Sanktionen: 5.000 Dollar und Briefe an den Mandanten und an jeden Richter, der als Verfasser einer Fälschung genannt war.

Auf die Frage „Ist Varghese ein echter Fall“ antwortete der Chatbot, er sei echt und in den juristischen Datenbanken zu finden. Das Gericht gewichtete das Festhalten an den Fälschungen nach den Warnungen stärker als den ersten Fehler.

UNSERE LESART

Das Werkzeug zu fragen, ob seine Quelle echt ist, ist keine Prüfung. Die Quelle zu öffnen, ist eine.

Quelle: United States District Court for the Southern District of New York, 2023

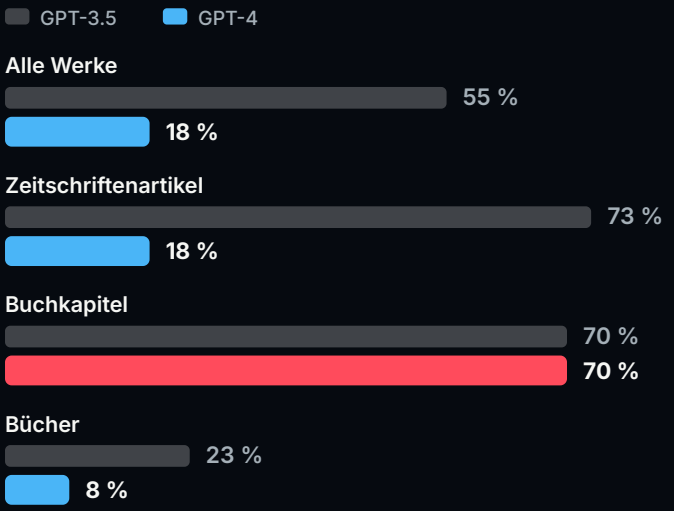
Aus der Entscheidung des US-Bezirksgerichts für den Southern District of New York vom 22. Juni 2023, vollständig gelesen; alle Daten 2023. Die Anwälte werden hier nicht genannt. Das Gericht stellte Bösgläubigkeit fest und dass auch die zitierten echten Fälle das Argument meist nicht stützten.

DIE ZAHLEN

# 636 Belege, einzeln geprüft

Im April 2023 ließen William Walters und Esther Isabelle Wilder GPT-3.5 und GPT-4 kurze Literaturübersichten mit Quellen zu 42 Themen schreiben, von E-Zigaretten bis Brexit, und suchten dann jedes der 636 zitierten Werke.

ZITIERTE WERKE, DIE ES NICHT GIBT, NACH ART, 84 KURZE ÜBERSICHTEN (2023)




Von den Werken, die es gibt, wurden 43 % (GPT-3.5) und 24 % (GPT-4) mit einem inhaltlichen Fehler zitiert, meist bei den Zahlen: falsches Jahr bei 22 % und 16 %, falscher Band, falsches Heft oder falsche Seiten bei 34 % und 13 % der Artikel und Kapitel. Fünf Übersichten gaben sich als Studien aus, mit erfundenen Methoden und Ergebnissen.

UNSERE LESART

Ein neueres Modell erfand weniger Quellen, nicht keine. Die Zahlen eines Belegs sind der Teil, den man Ziffer für Ziffer prüft.

Quelle: William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023

Walters & Wilder (2023), vollständig gelesen: Anteile von 222 (GPT-3.5) und 414 (GPT-4) zitierten Werken; ein Werk galt als echt, wenn Titel und Autoren ganz oder fast übereinstimmten. Ein Test der Versionen von 2023, nicht heutiger Werkzeuge.

DAS MODELL

# Falsch oder falsch belegt

Varun Magesh, Daniel Ho und Kollegen aus Stanford und Yale beurteilen eine KI-Antwort nach zwei Dingen: Ist sie richtig, und stützen ihre Quellen sie? Ein erfundener Fall ist das sichtbare Versagen. Ein echter Fall, zitiert für etwas, das er nicht sagt, ist schwerer zu sehen.

## Halluziniert

Falsch: jede unwahre Aussage

Falsch belegt: eine Quelle, die die Aussage nicht stützt

## Unvollständig

Verweigerung: keine oder eine unpassende Antwort

Unbelegt: richtig, aber ohne Quelle für die Kernaussagen

Gefragt nach bedeutenden Urteilen von Richter Luther A. Wilgarten, den es nie gab, lieferte ein Werkzeug einen echten Fall eines anderen Richters. Das Zitat war echt, die Antwort falsch. Die Autoren nennen solche Fehler potenziell gefährlicher als einen erfundenen Fall, „weil sie subtiler und schwerer zu erkennen sind“.

UNSERE LESART

Dass eine Quelle existiert, ist die erste Prüfung, nicht die letzte. Die zweite ist, ob sie sagt, was der Text behauptet.

Quelle: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024

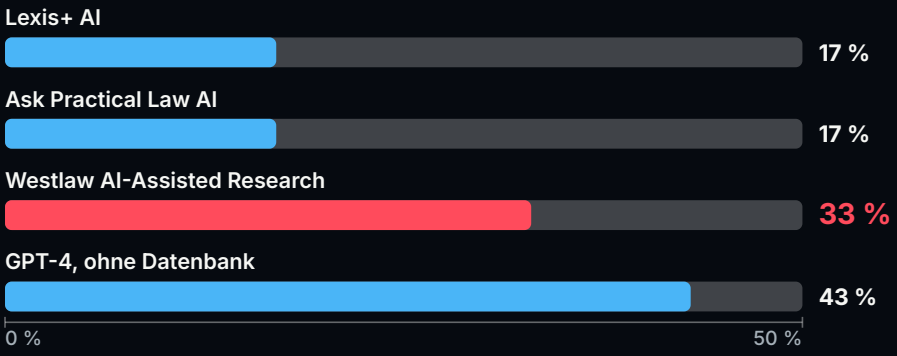
Definitionen nach Magesh et al. (Preprint, 2024). Teilweise richtige Antworten gelten als richtig; Belege werden nur bei richtigen Antworten beurteilt. Die zwei Spalten sind die Kategorien der Autoren, die Formulierung stammt von der Redaktion.

WAS DIE FORSCHUNG SAGT

# Für das Recht gebaut, trotzdem falsch

Anbieter sagten, ihre Rechtsrecherche-Werkzeuge, die aus eigenen Datenbanken antworten, vermieden oder beseitigten Halluzinationen. Im Frühjahr 2024 stellte dasselbe Team 202 schwierige, vorab registrierte Rechtsfragen an drei solche Werkzeuge und an GPT-4 und prüfte jede Antwort von Hand.

ANTWORTEN MIT HALLUZINATION, ANTEIL ALLER ANTWORTEN AUF 202 FRAGEN (2024)




Das beste Werkzeug antwortete in 65 % richtig und mit gültigen Quellen; ein anderes ließ 63 % unvollständig. Das Westlaw-Werkzeug, mit den meisten Fehlern der drei, schrieb auch die längsten Antworten, im Schnitt 350 Wörter gegen 219 und 175: mehr Aussagen, mehr zu prüfen. Das Fazit der Autoren: Nutzer „müssen weiterhin prüfen, ob die Kernaussagen durch die Zitate zutreffend gestützt sind“.

UNSERE LESART

Antworten aus einer Datenbank senkten die Fehler gegenüber einem allgemeinen Chatbot. Die Quelle lesen muss man trotzdem.

Quelle: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024

Magesh et al., Preprint vollständig gelesen; die Zeitschriftenfassung von 2025 nicht. Die Fragen waren bewusst schwer, die Anteile sind also keine Raten für den Alltag; Werkzeuge ändern sich. Verwendet wird die aktualisierte Grafik der Autoren.

WIE ES GEMESSEN WIRD

# Jede Quelle öffnen

Magesh und Kollegen beschreiben, was Prüfen heißt: zur Quelle durchklicken, sie lesen, ihre Autorität beurteilen, sie mit der Aussage vergleichen. Sechs Fragen machen daraus eine Zählung, die ein Team für jeden mit KI entworfenen Text führen kann.

| | | |
|---|---|---|
| 01 | **Es gibt sie** | die Quelle öffnet sich; das Wort des Werkzeugs zählt nicht |
| 02 | **Sie sagt es** | die Stelle stützt die Aussage, als Wort des Autors, nicht einer Partei |
| 03 | **Sie gilt noch** | nicht aufgehoben, ersetzt oder veraltet |
| 04 | **Sie passt hierher** | richtiger Ort, Zeitraum und Fall |
| 05 | **Die Zahlen stimmen** | Jahre, Seiten und Zahlen gegen das Original |
| 06 | **Es ist markiert** | was ein Mensch geprüft hat, getrennt vom Erzeugten |

UNSERE LESART

Zählen Sie die an einer geöffneten Quelle geprüften Aussagen, nicht die Minuten. Ein kurzer, ganz geprüfter Entwurf schlägt einen langen, überflogenen.

Quellen: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024; William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023; United States District Court for the Southern District of New York, 2023

Die Fragen stützen sich auf die Fehlerarten der Rechtsstudie, den Befund von Walters & Wilder, dass Zahlen am häufigsten falsch sind, und die Gerichtsentscheidung; die Auswahl stammt von der Redaktion. KI bei der Arbeit steht in Nr. 10, 51 und 86, Vertrauen in automatische Signale in Band 3 der Abhandlung.

WERKZEUG DER AUSGABE

# Die Karte vor dem Versand

Eine Karte für jeden Text, der das Team mit Maschinenhilfe verlässt: ein Bericht, eine Zusammenfassung, ein Brief. Das Werkzeug Ops Report schreibt den Periodenbericht aus den eigenen Protokollen des Teams; die Karte ist der Schritt zwischen Entwurf und Versand.

| | | |
|---|---|---|
| 01 | **Was rausgeht, an wen** | der Text, sein Leser, die Entscheidung, die er stützt |
| 02 | **Aussagen mit Quelle** | wie viele; jede wird geprüft, nicht vom Werkzeug übernommen |
| 03 | **Quellen geöffnet und gelesen** | n von N; eine, die sich nicht öffnet, lässt ihre Aussage ohne Quelle |
| 04 | **Zahlen am Original geprüft** | Daten, Beträge, Versionen, Seiten |
| 05 | **Vom Werkzeug ergänzt** | Namen, Uhrzeiten oder ein „erledigt", das nicht in den Notizen stand |
| 06 | **Geprüft von, und wann** | ein Name; wer unterschreibt, hat es gelesen |

Quellen: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024; William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023; United States District Court for the Southern District of New York, 2023

Die Karte ist ein Vorschlag der Redaktion, kein geprüftes Instrument; sie stützt sich auf die drei Quellen und den verlinkten Essay, der die eigene Regel des Autors für KI-Entwürfe schildert. Persönliche und vertrauliche Daten gehören nicht in nicht freigegebene Werkzeuge.

Mehr im Essay — Der Bericht, den eine Maschine geschrieben hat

Tool öffnen: Betriebsbericht — stivencatalyst.com/de/tools/ops-report/

QUELLEN UND METHODE

# Jede Zahl hat eine Quelle.

Die Zahlen dieser Ausgabe stammen aus den folgenden Quellen. Das Jahr zeigt, wie aktuell jede ist.

**Mata v. Avianca, Inc., No. 22-cv-1461 (PKC): Opinion and Order on Sanctions**
United States District Court for the Southern District of New York, 2023
https://storage.courtlistener.com/recap/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0.pdf

**Fabrication and errors in the bibliographic citations generated by ChatGPT**
William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023
https://doi.org/10.1038/s41598-023-41032-5

**Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools**
Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024
https://arxiv.org/abs/2405.20362

REDAKTIONELLE METHODE

Jede Zahl wurde auf Jahr, Herausgeber und Bedeutung geprüft, am Volltext, wo er sich öffnen ließ. Eine nur in einer Zusammenfassung, einem Abstract oder Auszügen gelesene Quelle ist mit „über“ markiert. Eine Zahl, die sich im Gelesenen nicht bestätigen ließ, ist auf ihrer Seite so markiert oder fehlt. Die Deutung des Herausgebers ist als „Unsere Lesart“ markiert.



# MANAGEMENT REVIEW

Management ohne Theater.

Jede Ausgabe eine Führungsfrage, geprüft an der besten Forschung.

## ALLE AUSGABEN

stivencatalyst.com/de/magazine/management-review.html

Management Review · Nr. 96 · November 2026 · Stiven Catalyst
Herausgegeben von Stiven Janaqi, der sie verantwortet; „die Redaktion" ist in dieser Publikation er. Texte, Übersetzungen, Layout und die Prüfung der Zahlen an ihren Quellen entstehen in Arbeitssitzungen mit Claude, einem KI-Modell von Anthropic.
stivencatalyst.com/de/how-this-is-made.html