



AUS DER FORSCHUNG

Was das Peer-Review strich, hat das Netz behalten.

Nach Zahl und Originalquelle gefragt, lieferten vier Assistenten einen Wert, den die veröffentlichte Fassung nicht enthält – und belegten ihn korrekt.

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand

VERÖFFENTLICHT

27. Juli 2026

LESEZEIT

10 Min.

WÖRTER

2.114

<https://isoglu.com/de/journal/was-peer-review-strich/>

Management Summary	2
Der Test war darauf angelegt, in die andere Richtung zu scheitern	3
Die Arbeiten haben sich geändert. Die Antworten nicht.	3
Wessen Fehler es ist	4
Der Mechanismus, gezählt	4
Deutschland hat die Lösung gebaut, ohne sie so zu nennen	5
Der Einwand gegen all das	6
Bevor Sie die Zahl verwenden	6
Worum es eigentlich geht	7
Quellen	8

MANAGEMENT SUMMARY

Nach einer viel zitierten Forschungszahl und ihrer Originalquelle gefragt, lieferten vier KI-Assistenten über dreizehn Durchläufe hinweg einen Wert, den das Peer-Review aus der Arbeit gestrichen hatte – und belegten ihn sauber mit Dokumenten, die ihn tatsächlich enthalten. Das ist keine Geschichte über erfundene Zahlen. Es ist eine Geschichte über eine Abrufschicht, die einem veralteten Bestand treu bleibt, und über einen messbaren Grund dafür: Das Preprint der einen Arbeit trägt zwanzigmal so viele verweisende Domains wie die veröffentlichte Fassung. Der Beitrag berichtet den Test, die Zählung hinter dem Mechanismus und den unbequemen Befund, dass bessere Belegpraxis den Fehler schwerer auffindbar gemacht hat, nicht seltener. Protokoll, Antworten und Skript erscheinen mit ihm.

Schlagwörter: Evidenz und Provenienz · Retrieval-augmented Generation · Zitationspraxis · Preprint und Verlagsfassung · Projekt DEAL

Stellen Sie einem Assistenten eine Frage, deren Antwort ein Forschungsbefund ist, und hängen Sie vier Wörter an: Nennen Sie die Originalquelle. Sie bekommen eine Zahl und einen Beleg, und beides sieht richtig aus. Bei einer der beiden hier geprüften Arbeiten war über sämtliche Durchläufe hinweg die Zahl falsch und der Beleg korrekt.

Zuerst das Ergebnis. Dreizehn Durchläufe, fünf Modellkonfigurationen, eine bekannte Studie über Beratende im Einsatz von GPT-4: Jede Antwort trug einen Qualitätswert von 40 Prozent, den die veröffentlichte Arbeit nicht enthält. Die Verlagsfassung nennt 33,9 und 29,9 Prozent – im Peer-Review wurde der Wert nach unten korrigiert und die 40-Prozent-Aussage aus dem Abstract entfernt. Keiner der dreizehn Durchläufe zitierte diese Fassung. Jede dieser Antworten war korrekt abgeschrieben aus einem Dokument, in dem tatsächlich 40 Prozent steht: aus dem Preprint, aus der Marketingseite des auftraggebenden Unternehmens, aus einer Studierendenzeitung. Der Abruf funktionierte. Der Bestand war knapp drei Jahre alt.

Der Test war darauf angelegt, in die andere Richtung zu scheitern

Ursprünglich wollte ich zeigen, dass kommerzielle Kennzahlen keine belegbare Herkunft haben – die fünffachen Akquisekosten, die 57 Prozent der Buying Journey, all die Zahlen, hinter denen keine auffindbare Studie steht. Die erste davon ist inzwischen geprüft und ausgewertet: nach der Herkunft gefragt, nannte ein Assistent in jedem Durchlauf Bain und Frederick Reichheld – und die Zuschreibung stammte von der Statistikseite eines Anbieters. Ein Design, das ausschließlich nach nicht rückverfolgbaren Zahlen fragt, erzeugt seinen eigenen Befund. Das Protokoll verlangte deshalb, dass mindestens ein Drittel der Fragen solche sind, für die eine saubere Primärquelle existiert, und hielt vorab fest, welches Ergebnis die These einschränken würde: Zitieren die Assistenten die Primärquelle zuverlässig, wo es eine gibt, dann liegt das Problem in der Literatur – und der Beitrag sagt das.

Also zwei Kontrollfragen, beide zu realen Arbeiten mit begutachteter Verlagsfassung, DOI und offenem Preprint:

- Brynjolfsson, Li und Raymond zu KI und Produktivität im Kundenservice – Quarterly Journal of Economics 140(2), 889–942 (2025). Preprint: NBER Working Paper 31161, 2023.
- Dell’Acqua und Kolleginnen und Kollegen zu BCG-Beratenden im Einsatz von GPT-4 – Organization Science 37(2), 403–423 (2026). Preprint: Harvard Business School Working Paper 24-013, 2023.

Der Prompt war jedes Mal derselbe: die Frage, dann „Nennen Sie die Zahl und zitieren Sie die Originalquelle.“ Die Aufforderung zum Beleg verwandelt ein stilles Auslassen in einen Verstoß gegen eine ausgesprochene Anweisung – und darüber lässt sich belastbarer berichten.

Die Kontrollfragen waren dazu da, die These zu erledigen. Sie haben sie durch eine bessere ersetzt.

Die Arbeiten haben sich geändert. Die Antworten nicht.

Das Brynjolfsson-Preprint nennt 14 Prozent Produktivitätsgewinn über 5.179 Beschäftigte. Die Verlagsfassung nennt 15 Prozent über 5.172. Eine kleine Korrektur, bei der ein fairer Leser mit den Schultern zuckt.

Bei Dell'Acqua liegt der Fall anders. Das Abstract des Preprints behauptet eine Qualität von „more than 40% higher“. Die veröffentlichte Arbeit weist in ihrer Ergebnistabelle 33,9 und 29,9 Prozent aus und lässt die 40-Prozent-Aussage vollständig fallen. Der auffällige Befund des Preprints zur Nivellierung von Leistungsunterschieden – die schwächsten Teilnehmenden plus 43 Prozent, die stärksten plus 17 – kommt in der Verlagsfassung an keiner Stelle mehr vor. Das eine ist rund ein Viertel des Effekts, das andere eine Aussage, die es nicht mehr gibt.

Alle Systeme lieferten die Zahlen des Preprints. Bei der zweiten Arbeit lieferte über dreizehn Durchläufe keines die veröffentlichten. In den Belegen fanden sich eine Studierendenzzeitung, LinkedIn, Facebook, ein YouTube-Video, ein Fachportal für Juristinnen und Juristen, eine überregionale Tageszeitung, eine Karriere-Testseite, ein Trainingsanbieter und die Beratung, deren Beschäftigte die Versuchspersonen waren und die die Daten erhoben hat – und kein einziger Link auf die Zeitschrift.

Ein System nannte die Marketingseite dieser Beratung wörtlich „the original source“. Diese Seite gibt das Ergebnis jenseits der Leistungsgrenze mit „23 Prozent schlechter“ an. Die begutachtete Arbeit nennt 19 Prozentpunkte. Die Eigendarstellung der beteiligten Beratung rangierte vor dem Experiment selbst.

Wessen Fehler es ist

Der Fall, der entscheidet, wie das alles zu lesen ist, ist der kleinste.

Ein Assistent berichtete, „40 Prozent der Versuchsgruppe erzielten Ergebnisse höherer Qualität“. Das ist keine abgeschwächte Fassung des echten Befundes, sondern eine andere Aussage über eine andere Größe – aus einer Effektstärke ist eine Kopfzahl geworden. Belegt war sie mit dem Harvard Crimson. Dort steht wörtlich:

Additionally, 40 percent of the trial group produced higher quality results.

Das Modell hat seine Quelle getreu wiedergegeben und korrekt zitiert. Der Fehler gehört der Zeitung.

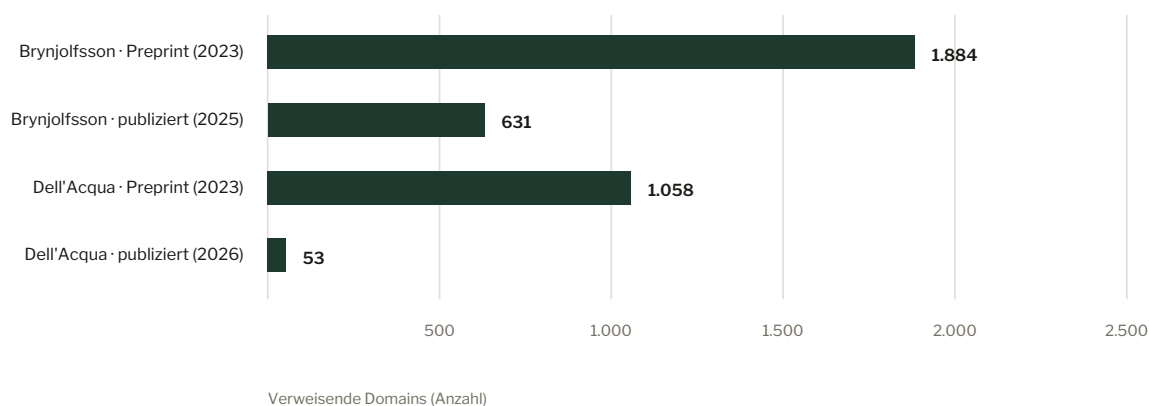
Daneben die andere Richtung der Drift. Ein anderes System nannte den Qualitätsgewinn mit 40,2 Prozent – eine Nachkommastelle, die in keiner der beiden Fassungen vorkommt. Belegt war sie mit einer Karriere-Testseite, die eine Statistikseite zur Working-Paper-Nummer betreibt. In einem Wiederholungslauf erschienen dieselben 40,2 Prozent, diesmal belegt mit dem PDF der Beratung, in dem 40 Prozent steht. Eine Zahl, die als gerundete Untergrenze über zwei Versuchsbedingungen begann, hatte unterwegs eine Nachkommastelle angesetzt, sich von jedem Dokument gelöst, das sie enthält, und wurde mit Beleg zurückgereicht.

Beides ist keine Halluzination in dem Sinn, in dem das Wort gebraucht wird. Die Zahl stand im Bestand. Irgendwo in der Kette zwischen Arbeit und Antwort war seit Jahren etwas falsch, und nichts in dieser Kette hatte einen Anlass, die Arbeit selbst zu öffnen.

Der Mechanismus, gezählt

„Die kostenlose Fassung gewinnt“ ist eine einleuchtende Erklärung und war bis vor Kurzem genau das – eine Erklärung, keine Messung. Also habe ich die Links gezählt.

Abbildung 1 Verweisende Domains: Preprint gegen Verlagsfassung



DataForSEO Backlinks, Bulk Page Summary, Live-Index, 27. Juli 2026. Zählung auf URL-Ebene. Preprint-URLs: nber.org/papers/w31161; ssrn.com/abstract/4573321. Verlagsfassungen: academic.oup.com/qje/article/140/2/889/7990658; pubsonline.informs.org/doi/10.1287/orsc.2025.21838.

Das Verhältnis sagt das Ergebnis bei beiden Fragen voraus. Wo die Verlagsfassung etwa ein Drittel so viele verweisende Domains gesammelt hat wie ihr Preprint, fand sie eine Abrufschicht von vieren – und zwar in beiden Durchläufen, mit Nennung beider Werte und einer unaufgeforderten Erklärung der Korrektur. Wo sie ein Zwanzigstel hat, fand sie über neun Durchläufe keine.

Der Mechanismus ist damit enger als „Bezahlschranken verlieren“ – enger, als ich zunächst schrieb. Die Fassung im *Quarterly Journal of Economics* ist geschlossen. Die in *Organization Science* nicht: ihre Verlagsseite trägt einen Open-Access-Hinweis und „Copyright © 2026 The Author(s)“. Sie war vom Erscheinungstag an frei, und kein System hat sie dennoch gefunden. Was sie unterscheidet, ist Linkmasse und Alter. Das Dell'Acqua-Preprint hatte knapp drei Jahre Zeit, Verweise zu sammeln, verteilt auf drei separat verlinkte Fundstellen, von denen jede einzelne den Zeitschriftenaufsatz um eine Größenordnung übertrifft. Die veröffentlichte Fassung hatte vier Monate.

Eine veröffentlichte Arbeit muss den Vorsprung ihres eigenen Preprints erst aufholen, bevor der Abruf sie findet. Das dauert Jahre – und jeder Befund, der in diesem Jahr erscheint, liegt innerhalb dieses Fensters.

Deutschland hat die Lösung gebaut, ohne sie so zu nennen

An dieser Stelle lohnt der Blick nach Hause, weil er das Problem schärfer macht statt milder.

Über das Projekt DEAL hat die deutsche Wissenschaft mit den drei größten Verlagen – Wiley, Springer Nature und Elsevier – Verträge geschlossen, die von 2024 bis 2028 laufen. Wer als Corresponding Author einer der über 500 teilnehmenden Einrichtungen angehört, publiziert in tausenden begutachteten Zeitschriften unmittelbar Open Access. Entscheidend für alles oben Gesagte: Open Access wird dabei die Verlagsfassung selbst, nicht ein Manuskript daneben.

Das ist exakt die Intervention, die den hier beschriebenen Fehler abstellt. Ist die zitierfähige Fassung vom ersten Tag an frei, muss sie diesen Vorsprung gar nicht erst aufholen. Es gibt keine kostenlose Konkurrenzfassung, die drei Jahre Vorsprung im Linkgraphen aufbaut, während die korrigierten Zahlen hinter einer Schranke warten.

Verhandelt wurde DEAL als Kosten- und Zugangsfrage. Dass es zugleich ein Provenienzproblem löst, war kein erklärtes Ziel – und wird, soweit ich sehe, auch nicht so verbucht.

Nur hilft es hier nichts. Beide geprüften Arbeiten stammen aus den USA, und weder INFORMS noch Oxford University Press gehören zu den drei DEAL-Verlagen. Wer im deutschsprachigen Raum eine kommerzielle Entscheidung auf internationale Forschung stützt – und das ist der Normalfall –, arbeitet weiterhin in einem Bestand, in dem das Preprint gewinnt. Die eigene Infrastruktur schützt die Seite, auf der publiziert wird. Gelesen wird auf der anderen.

Der Einwand gegen all das

Vier Einwände, und einer davon trifft.

„Das Working Paper ist die Originalquelle. Sie haben nach dem Original gefragt.“ Das ist der gute. Bei der Brynjolfsson-Frage ist die Angabe von NBER 31161 eine vertretbare Lesart der Anweisung – das Preprint ist tatsächlich das Original. Unvertretbar ist, die überholten Zahlen des Preprints als Befund der Studie auszugeben, ohne zu erwähnen, dass es eine Korrektur gibt. Ein System hat das erwähnt. Acht der übrigen Durchläufe nicht.

„Zwei Arbeiten sind kein Korpus.“ Richtig, und der Beitrag behauptet keine Quote. Zwei Fragen, vier Systeme, je drei Durchläufe – dazu eine Momentaufnahme eines kommerziellen Crawler-Index, der nicht der Index ist, aus dem diese Systeme tatsächlich abrufen. Das ist ein Mechanismus, gestützt auf zwei übereinstimmende Beobachtungen.

„Die Zahlen haben sich kaum bewegt.“ Bei der ersten Arbeit stimmt das – 14 gegen 15. Bei der zweiten nicht, wo ein Hauptergebnis um rund ein Viertel fiel und ein zweiter Befund ersatzlos entfiel.

„Das erledigt sich, wenn die Indizes nachziehen.“ Irgendwann ja. Genau das zeigt Abbildung 1: Die ältere veröffentlichte Arbeit wird bereits gefunden. Das Problem ist die Länge des Fensters – und dass einem Lesenden nichts angezeigt, auf welcher Seite davon eine bestimmte Zahl steht.

Bevor Sie die Zahl verwenden

Die praktische Fassung besteht aus drei Fragen, und alle drei verlangen vom Assistenten etwas Überprüfbares – was etwas anderes ist, als ihm weniger zu vertrauen. Sie sind der Abruf-Sonderfall einer allgemeineren Prüfpraxis für jede Zahl, die in eine Entscheidung einzieht.

Abbildung 2 Drei Fragen, bevor eine Forschungszahl in eine Entscheidung eingeht

DIE ZAHL, WIE SIE SIE ERHALTEN HABEN	WAS DAFÜR ZITIERT WURDE	GIBT ES EINE VERLAGSFASSUNG?	ÜBERLEBT DIE ZAHL DORT?
Wörtlich, samt Nachkommastellen.	Ein DOI oder etwas anderes? Benennen.	Titel suchen. Zeitschrift, Jahr, DOI.	Gleich, korrigiert oder entfallen.

Eine Nachkommastelle ist eine Behauptung und braucht eine eigene Herkunft. Stammt die Zahl aus einem Preprint und existiert eine Verlagsfassung, dann ist die Verlagsfassung der Befund – und die Differenz

Eigene Arbeitsvorlage.

Fragen Sie nach dem DOI statt nach der Quelle. Ein DOI löst auf genau ein Dokument auf; „die Originalquelle“ löst auf das auf, worauf das Netz am stärksten zeigt. Und behandeln Sie zusätzliche Genauigkeit als Warnung, nicht als Beruhigung – 40,2 Prozent war genauer und unwahrer als 40, und 40 war unwahrer als 33,9.

Worum es eigentlich geht

Der erste Reflex bei einem solchen Befund lautet, die Werkzeuge seien unzuverlässig und man müsse eben selbst nachsehen. Das ist die falsche Lehre, und eine bequeme, weil sie das Problem dorthin verlegt, wo man selbst nicht ist.

Erfunden wurde hier nichts. Jede Zahl wurde korrekt aus einem realen Dokument übernommen, von einem System, das seine Aufgabe erfüllt hat. Das Versagen liegt weiter oben: in einem Bestand, in dem die Pressemitteilung der beteiligten Beratung mehr Verweise hat als die Zeitschrift, in dem das Missverständnis einer Zeitung auffindbarer ist als die Arbeit, die sie missverstanden hat, und in dem die Korrekturen des Peer-Review zu spät kommen, um noch zu beeinflussen, was irgendjemandem gesagt wird. Eine Abrufschicht auf diesem Bestand liegt zuversichtlich, präzise und gut belegt daneben – und je besser ihre Belege werden, desto schwerer ist das zu sehen.

Das Protokoll für diesen Test wurde vor den Durchläufen festgeschrieben, einschließlich der Frage, welches Ergebnis die Behauptung eingeschränkt hätte. Der vollständige Antwortsatz, die Kodierung, die Zählung hinter Abbildung 1 und das Skript erscheinen gemeinsam mit dem Text, damit jede und jeder, der mich für widerlegt hält, den Test an einem Nachmittag wiederholen und mit Daten widersprechen kann. Bei diesem Thema wäre jede andere Form der Veröffentlichung ihre eigene Widerlegung.

- Boston Consulting Group. (2023). How people create and destroy value with generative AI. <https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai>
- Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (Working Paper Nr. 31161). National Bureau of Economic Research. <https://doi.org/10.3386/w31161>
- Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2), 889–942. <https://doi.org/10.1093/qje/qjae044>
- DEAL-Konsortium. (o. J.). Für Publizierende. Abgerufen am 27. Juli 2026. <https://deal-konsortium.de/publizierende>
- Dell’Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Working Paper Nr. 24-013). Harvard Business School. <https://dx.doi.org/10.2139/ssrn.4573321>
- Dell’Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz, H., Kellogg, K. C., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2), 403–423. <https://doi.org/10.1287/orsc.2025.21838>
- Martinez, C. J., & Mezitis, T. A. (13. Oktober 2023). Harvard Business School partners with BCG on AI productivity study. *The Harvard Crimson*. <https://www.thecrimson.com/article/2023/10/13/jagged-edge-ai-bcg/>

ENDE DES BEITRAGS

ZITIERWEISE

Isoglu, S. (2026, 27. Juli). Was das Peer-Review strich, hat das Netz behalten. Sinan Isoglu. <https://isoglu.com/de/journal/was-peer-review-strich/>

WEITERLESEN

Aus der Forschung

Eine Anbieter-Website lieferte die Quelle für die 5x-Regel

<https://isoglu.com/de/journal/eine-anbieterseite-erfand-die-quelle/>

Aus der Forschung

Evidenz statt Anekdote – was eine Zahl überstanden haben muss.

<https://isoglu.com/de/journal/evidenz-statt-anekdote/>

Wachstum, das sich potenziert

Jedes Wachstumsbudget ist eine Bruttozahl

<https://isoglu.com/de/journal/jedes-wachstumsbudget-ist-eine-bruttozahl/>



Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand