KONZEPTIONELLER BEITRAG · LITERATURBASIS 2015–2026

Schema, Entitäten und die Grenzen maschinenlesbarer Auszeichnung

Was generative Antwortsysteme von einer Website tatsächlich lesen

Von Victor Frenzel, KI-Sichtbarkeit und GEO

VICTORFRENZEL.COM · KÖLN · AUGUST 2026 · KONZEPTIONELLER BEITRAG AUF LITERATURBASIS

Abstract

In der Optimierungspraxis hat sich eine Kausalkette etabliert, die selten ausgesprochen und noch seltener geprüft wird: Strukturierte Auszeichnung nach dem Schema.org-Vokabular lasse ein generatives Antwortsystem ein Unternehmen als Entität erkennen, und diese Erkennung erhöhe die Wahrscheinlichkeit, in einer synthetischen Antwort genannt zu werden. Der vorliegende Beitrag nennt diese Kette die Auszeichnungsprämisse und prüft ihre drei Glieder einzeln gegen den verfügbaren Forschungsstand. Das Ergebnis ist asymmetrisch. Für das erste Glied, die Lesbarkeit, liegt Evidenz vor, dass mehrere Antwortsysteme eingebettetes JSON-LD im Abrufpfad nicht zurückgeben (Tousseyn, 2026). Für das zweite Glied, die Notwendigkeit der Auszeichnung zum Verstehen, spricht die Architektur retrieval-gestützter Generierung, die auf Textrepräsentationen operiert (Lewis et al., 2020). Für das dritte Glied, die Wirkung auf die Nennung, findet sich in der einschlägigen Experimentalliteratur kein Befund, weil Auszeichnung dort in keinem der geprüften Merkmalssätze vorkommt (Aggarwal et al., 2024; Vishwakarma et al., 2026; Zhang et al., 2026). Eine kritische Übersicht über fünfundvierzig Arbeiten hält fest, dass keine geprüfte Technik einen stabilen, längsschnittlichen, plattformübergreifenden kausalen Effekt auf organische Auffindbarkeit zeigt (Martinez, 2026). Die Betreiberdokumentation des größten Anbieters widerspricht der Prämisse ausdrücklich (Google, 2026). Ein Gegenbefund verlangt jedoch eine Verengung: Auf Googles eigenen Oberflächen stieg die Zitationshäufigkeit nach einer Schema-Einführung erheblich, was die Studienautoren selbst auf den klassischen Suchpfad und teilweise auf einen Algorithmuswechsel zurückführen (Tousseyn, 2026). Der Beitrag schlägt daher vor, Auszeichnung als Hebel der klassischen Suche zu führen, nicht als Hebel generativer Sichtbarkeit, und Entitätsklarheit von ihrer technischen Notation zu trennen.

Schlüsselwörter: Generative Engine Optimization, strukturierte Daten, Schema.org, Entitäten, KI-Sichtbarkeit, Zitationsauswahl

English Abstract

Optimization practice has settled on a causal chain that is rarely stated and even more rarely tested: structured annotation using the Schema.org vocabulary supposedly enables a generative answer system to recognise a business as an entity, and this recognition supposedly raises the probability of being named in a synthetic answer. This paper calls that chain the annotation premise and examines its three links separately against the available research. The result is asymmetric. For the first link, readability, evidence indicates that several answer systems do not return embedded JSON-LD along the

retrieval path (Tousseyn, 2026). For the second link, the necessity of annotation for comprehension, the architecture of retrieval-augmented generation, which operates on text representations, argues against it (Lewis et al., 2020). For the third link, the effect on being named, the relevant experimental literature contains no finding at all, because annotation appears in none of the tested feature sets (Aggarwal et al., 2024; Vishwakarma et al., 2026; Zhang et al., 2026). A critical survey of forty-five studies states that no reviewed technique demonstrates a stable, longitudinal, cross-platform causal effect on organic discoverability (Martinez, 2026). The operator documentation of the largest provider contradicts the premise explicitly (Google, 2026). One counter-finding requires a narrowing, however: on Google's own surfaces, citation frequency rose substantially after a schema rollout, which the study's own authors attribute to the classical search path and partly to an algorithmic shift (Tousseyn, 2026). The paper therefore proposes treating annotation as a lever of classical search rather than of generative visibility, and separating entity clarity from its technical notation.

Keywords: generative engine optimization, structured data, schema.org, entities, AI visibility, citation selection

1. Einleitung

Die Praxisliteratur zur Sichtbarkeit in generativen Antwortsystemen hat in kurzer Zeit ein festes Erzählmuster ausgebildet. Es lautet ungefähr so: Suche habe sich von Zeichenketten zu Dingen und von Dingen zu Entitäten verschoben, ein Antwortsystem baue sich zuerst ein internes Modell des Unternehmens und empfehle danach nur, was es zuvor eindeutig auflösen könne. Aus dieser Beschreibung folgt eine Handlungsempfehlung, die technisch anspruchsvoll klingt und gut verkäuflich ist: strukturierte Auszeichnung nach dem Schema.org-Vokabular, dauerhafte Kennungen für jede Entität, serverseitige Ausspielung, kontinuierliche Überwachung der Auszeichnungsschicht.

Der vorliegende Beitrag nimmt dieses Muster ernst genug, um es zu prüfen. Anlass war ein Agenturbeitrag, der die Systematik in drei Stufen ordnet und im zweiten Teil das eigene Produkt bewirbt. Solche Texte sind nicht zitierfähig, sie sind aber ein zuverlässiger Indikator dafür, welche Annahmen im Markt gerade als selbstverständlich gelten. Die hier verfolgte Systematik ist eine andere und wird eigenständig begründet.

Der Beitrag verfolgt drei Ziele. Erstens wird die unterstellte Kausalkette explizit gemacht und in drei prüfbare Glieder zerlegt. Zweitens wird jedes Glied einzeln gegen die verfügbare Evidenz geprüft, mit ausdrücklicher Unterscheidung nach Belegklassen. Drittens wird aus dem Ergebnis eine Empfehlung abgeleitet, die weniger verspricht als die Praxisliteratur, dafür aber trägt.

2. Problemstellung und Erkenntnisinteresse

Generative Antwortsysteme beantworten eine Anfrage, indem sie Quellen abrufen und deren Inhalt zu einer Antwort verdichten. Aggarwal et al. (2024) beschreiben sie als Systeme, die Anfragen typischerweise dadurch beantworten, dass sie Informationen aus mehreren Quellen synthetisieren und mithilfe großer Sprachmodelle zusammenfassen, und halten fest, dass Inhalteanbieter angesichts der Intransparenz dieser Systeme kaum Kontrolle darüber besitzen, wann und wie ihre Inhalte dargestellt werden. Genau diese Intransparenz erzeugt den Markt, in dem die Auszeichnungsprämisse gedeiht. Wo nicht beobachtbar ist, was wirkt, wird plausibel, was vertraut ist.

Vertraut ist die Auszeichnung deshalb, weil sie in der klassischen Suche eine belegte Funktion besitzt. Sie ist dort seit über einem Jahrzehnt die Eintrittskarte zu angereicherten Ergebnisdarstellungen. Der Schluss von dieser Funktion auf eine Funktion in generativen Systemen ist jedoch kein Schluss, sondern eine Übertragung, und Übertragungen brauchen eine Begründung.

Das Erkenntnisinteresse ist daher nicht, ob strukturierte Daten nützlich sind. Sie sind es, und dieser Beitrag empfiehlt an keiner Stelle, sie zu entfernen. Das Erkenntnisinteresse gilt der Kausalkette, die in der Praxisliteratur unterstellt wird, und der Frage, welches ihrer Glieder belegt ist und welches nicht.

3. Die Auszeichnungsprämisse

3.1 Begriffsbestimmung

Der Beitrag führt für die geprüfte Annahme den Begriff der **Auszeichnungsprämisse** ein. Gemeint ist die Annahme, dass die Anwesenheit maschinenlesbarer Auszeichnung auf einer Seite die Wahrscheinlichkeit erhöht, dass ein generatives Antwortsystem die beschriebene Organisation in einer synthetischen Antwort nennt. Der Begriff ist in dieser Fassung neu und wird als analytisches Werkzeug vorgeschlagen, nicht als empirisch geprüftes Konstrukt. Er ist ausdrücklich eine eigene Begriffsprägung des Verfassers.

Die Prämisse zerfällt in drei Glieder, die in der Praxisliteratur meist zu einem einzigen Satz zusammengezogen werden:

1. **Lesbarkeit.** Das Antwortsystem nimmt die Auszeichnung auf seinem Abrufpfad überhaupt auf.
2. **Notwendigkeit.** Das System benötigt die Auszeichnung, um den Inhalt und die beschriebene Organisation zu verstehen.

3. **Wirkung.** Das Verstehen über die Auszeichnung verschiebt die Auswahlentscheidung zugunsten dieser Quelle.

Nur wenn alle drei Glieder halten, trägt die Praxisempfehlung. Bricht eines, bleibt die Auszeichnung eine gute Idee aus anderen Gründen, aber kein Hebel generativer Sichtbarkeit.

3.2 Wofür Schema.org gebaut wurde

Für die Prüfung ist entscheidend, dass Schema.org ein Werkzeug mit dokumentiertem Entstehungszweck ist und nicht ein neutraler Beschreibungsstandard. Guha, Brickley und Macbeth (2016), die das Vokabular mit auf den Weg gebracht haben, beschreiben die Vorgeschichte offen: Weil die Suchmaschinen zunächst je eigene Vokabulare empfahlen, hätten die meisten Betreiber am Ende gar keine Auszeichnung ergänzt, und die ergänzte sei häufig falsch formatiert gewesen. Das gemeinsame Vokabular sollte diese Verwirrung beenden. Die Verfasser sind zugleich Mitarbeitende der beteiligten Anbieter, Guha und Brickley bei Google, Macbeth bei Microsoft. Ihr Bericht ist damit kein unabhängiger Drittbefund, sondern eine Selbstauskunft der Urheber, und wird hier so behandelt.

Zwei Designentscheidungen aus diesem Bericht sind für die Entitätsdebatte unmittelbar relevant. Zum einen verzichtet Schema.org bewusst auf global eindeutige Kennungen für die beschriebenen Dinge. Guha et al. (2016) begründen dies damit, dass die Abstimmung von Entitätsreferenzen mit anderen Seiten für die Zehntausende von Entitäten, über die eine Website Informationen besitzen könne, für die meisten Seiten schlicht zu schwierig sei. Zum anderen berichten sie über die Eigenschaft, die diese Abstimmung ermöglichen sollte, ernüchtert: Schema.org stelle zwar eine `sameAs`-Eigenschaft bereit, um Entitäten mit bekannten Seiten wie Startseiten oder Wikipedia zu verknüpfen und so die Zusammenführung zu erleichtern, doch habe diese keine nennenswerte Verbreitung gefunden.

Damit liegt eine erste Spannung offen. Die heutige Praxisliteratur verkauft als Kern ihrer Methode genau jene Entitätsverknüpfung, die die Urheber des Vokabulars als für die meisten Seiten zu aufwendig eingeschätzt und deren Umsetzung sie als kaum angenommen beschrieben haben.

4. Erstes Glied: Wird die Auszeichnung gelesen?

Die Verbreitung der Auszeichnung ist gut dokumentiert. Guha et al. (2016) berichteten für eine Stichprobe von zehn Milliarden Seiten einen Anteil von 31,3 Prozent mit Schema.org-Auszeichnung, gegenüber 22 Prozent ein Jahr zuvor. Brinkmann, Primpeli und Bizer (2023) führen die Reihe fort und weisen für den Common Crawl einen Anstieg von

3,1 Prozent aller erfassten Domains im Jahr 2013 auf 37,9 Prozent im Jahr 2022 aus, absolut von 400.000 auf über 12 Millionen Domains. Sie merken an, dass der Common Crawl nur eine Stichprobe von rund 30 Millionen populären Domains abdeckt, während das vollständige Web mehr Websites enthält.

Verbreitung ist jedoch nicht Rezeption. Die Frage, ob ein generatives Antwortsystem die Auszeichnung auf seinem Abrufpfad überhaupt aufnimmt, ist eine andere und schlechter belegte.

Der bislang direkteste Test stammt aus einer Anbieterstudie. Tousseyn (2026) führte zwischen dem 7. Dezember 2025 und dem 7. März 2026 auf der eigenen Unternehmenswebsite fünf Auszeichnungstypen ein und forderte anschließend sieben Antwortsysteme auf, den Auszeichnungscode der betreffenden Seiten zurückzugeben. Sechs von sieben Systemen konnten die Auszeichnung nicht abrufen oder nicht korrekt interpretieren; einzig Gemini gab das eingebettete JSON-LD zutreffend zurück. Ein System lieferte einen Auszeichnungstyp zurück, der auf der Seite nicht implementiert war. Der Studienautor kommentiert diesen Fall mit dem Satz, eine Plattform, die zuversichtlich falsche strukturierte Daten zurückgebe, lese kein Schema, sondern erzeuge Text, der wie Schema klinge. In einem dritten Testteil hinterlegte er eine Information ausschließlich in der Auszeichnung, ohne sie im sichtbaren Seitentext zu wiederholen. Kein System nutzte sie zur Beantwortung der zugehörigen Frage.

Diese Befunde sind vorsichtig zu gewichten. Es handelt sich um eine Anbieterstudie auf einer einzelnen Domain eines Softwareunternehmens, deren Autor die Übertragbarkeit auf andere Branchen ausdrücklich einschränkt. Der Fetch-Test misst zudem nicht den regulären Abrufpfad, sondern eine explizite Nutzeranweisung; ob ein System eine Auszeichnung im Hintergrund verarbeitet, die es auf Aufforderung nicht wiedergibt, bleibt offen. Der dritte Testteil ist von dieser Einschränkung weniger betroffen, weil er nicht das Zurückgeben, sondern das Verwerten prüft.

Ein zweiter Hinweis kommt aus der Betreiberdokumentation. Google (2026) beschreibt in der Anleitung zu agentischen Erfahrungen, wie Browser-Agenten Websites erschließen, und nennt dafür die Analyse visueller Darstellungen wie Bildschirmaufnahmen, die Inspektion der DOM-Struktur und die Interpretation des Barrierefreiheitsbaums. Strukturierte Auszeichnung kommt in dieser Aufzählung nicht vor. Das ist kein Beleg für Unwirksamkeit, wohl aber ein Hinweis darauf, dass der Betreiber selbst den agentischen Zugriff nicht über die Auszeichnungsschicht denkt.

5. Zweites Glied: Braucht ein Sprachmodell die Auszeichnung?

Das zweite Glied ist das architektonisch schwächste. Lewis et al. (2020) haben retrieval-gestützte Generierung als Verbindung eines vortrainierten Sequenzmodells mit einem nicht-parametrischen Gedächtnis eingeführt, das als dichter Vektorindex über einem Textkorpus realisiert und über einen neuronalen Retriever angesprochen wird. Der Abruf operiert damit auf Repräsentationen von Text, nicht auf typisierten Aussagen. Google (2026) beschreibt seine eigenen generativen Funktionen in denselben Begriffen und definiert retrieval-gestützte Generierung als Technik, die auf den Kernsystemen der Suche aufsetzt, um relevante Seiten aus dem Suchindex abzurufen und daraus eine Antwort zu erzeugen.

Damit kehrt sich die Beweislast um. Schema.org wurde für Parser gebaut, die unstrukturierten Text nicht zuverlässig deuten konnten; das ist der von Guha et al. (2016) berichtete Entstehungsgrund. Ein Sprachmodell hat diese Beschränkung nicht. Es benötigt keine Typangabe, um zu erkennen, dass ein Abschnitt Fragen und Antworten enthält, und keine Preisauszeichnung, um einen im Fließtext genannten Preis als Preis zu lesen. Wer behauptet, die Auszeichnung sei dennoch nötig, muss angeben, welche Information sie trägt, die der sichtbare Text nicht trägt. Genau dieser Fall wurde von Tousseyn (2026) im dritten Testteil konstruiert, und er ging negativ aus.

Hinzu kommt ein Qualitätsargument, das älter ist als die generative Suche. Meusel und Paulheim (2015) haben ausgelieferte Microdata-Auszeichnung über mehr als 250 Millionen Seiten quantitativ auf Fehler untersucht. Von den Domains, die Objekteigenschaften verwenden, belegen 56,58 Prozent mindestens eine davon mit einem Literalwert statt mit einem Objekt; bei 9,63 Prozent der Domains ließ sich mindestens ein Datentypwert nicht in den vorgesehenen Typ überführen. Die Verfasser halten den Umstand für so strukturell, dass sie eine konsumentenseitige Reparatur vorschlagen, mit der Begründung, es sei unrealistisch, dass Datenanbieter saubere und korrekte Daten liefern werden. Ihr Fazit fällt gleichwohl nicht vernichtend aus: Der überwiegende Teil der ausgezeichneten Information lasse sich gemäß dem empfohlenen Schema verarbeiten.

Alrashed, Paparas, Benjelloun, Sheng und Noy (2021) zeigen für einen einzelnen Typ, wie weit die Abweichung reichen kann. Beim Aufbau von Googles Datensatzsuche stellten sie fest, dass Seiten von 61 Prozent der Internet-Hosts, die eine Datensatz-Auszeichnung bereitstellen, tatsächlich gar keine Datensätze beschreiben. Ihre Antwort war kein Appell an die Betreiber, sondern ein tiefes neuronales Netz, das auszeichnungsführende Seiten klassifiziert und dabei 96,7 Prozent Trefferquote bei 95 Prozent Genauigkeit erreicht. Auch Guha et al. (2016) hatten diese Richtung bereits vorgezeichnet, als sie festhielten, Schema.org unterscheide sich von den üblichen Linked-Data-Leitlinien in

der Annahme, dass vor der Nutzung strukturierter Webdaten in Anwendungen üblicherweise Bereinigung, Abgleich und Nachbearbeitung nötig seien.

Die Konsequenz ist unbequem für die Praxisliteratur. Wenn große Konsumenten strukturierter Daten seit Jahren Klassifikatoren betreiben, um die Auszeichnung gegen den Seiteninhalt zu prüfen, dann ist der Seiteninhalt die maßgebliche Instanz und die Auszeichnung ein Hinweis, der validiert wird. Google (2025) formuliert genau diese Rangfolge als Empfehlung, wenn es darum bittet, sicherzustellen, dass die strukturierten Daten zum sichtbaren Text der Seite passen.

6. Drittes Glied: Was bewegt die Nennung tatsächlich?

Das dritte Glied lässt sich am schärfsten prüfen, weil hierzu kontrollierte Experimente vorliegen. Ihr gemeinsames Ergebnis ist bemerkenswert: Auszeichnung kommt in ihnen nicht vor.

Aggarwal et al. (2024) prüfen neun Optimierungsmethoden gegen einen Referenzkorpus. Die Liste umfasst autoritativen Stil, das Ergänzen von Statistiken, Keyword-Häufung, das Ergänzen von Quellenangaben und von Zitaten, Vereinfachung der Sprache, Verbesserung der Lesbarkeit sowie das Ergänzen seltener und fachsprachlicher Begriffe. Am stärksten wirken das Ergänzen von Quellenangaben, Zitaten und Statistiken mit relativen Verbesserungen von 30 bis 40 Prozent auf der positionsgewichteten Wortzahl und 15 bis 30 Prozent auf dem subjektiven Eindruck. Die aus der klassischen Optimierung übernommene Keyword-Häufung schneidet schlecht ab. Sämtliche neun Methoden operieren auf der Inhaltsebene; keine betrifft die Auszeichnungsschicht. Dass sie fehlt, ist keine Aussage der Verfasser über ihre Wirkungslosigkeit, sondern eine Beobachtung über den gewählten Untersuchungsraum, und wird hier ausdrücklich als solche geführt.

Vishwakarma, Kumar und Jamidar (2026) verengen die Frage auf den Wettbewerbsfall. In einem faktoriellen Aufbau werden je zwei Quellen in denselben Kontext eingespeist, die sich in genau einem von achtzehn Inhaltsmerkmalen unterscheiden; Markennamen sind anonymisiert, die Reihenfolge ist ausbalanciert. Über sechs Sprachmodelle hinweg werden 252.000 Durchgänge ausgewertet. Vier Merkmale wirken über alle sechs Modelle hinweg als Ausschlusskriterien: thematische Fehlpassung, fehlende Preisangabe, ein veralteter gegenüber einem aktuellen Zeitstempel und die schlechtere Listenposition. Sieben von achtzehn Merkmalen zeigen keinen belastbaren Effekt, darunter ausdrücklich die Gestaltungsmerkmale: Die Verfasser berichten, dass Formatierungsentscheidungen keinen Einfluss hatten, was nahelege, dass Sprachmodelle Inhalt unabhängig von seiner visuellen Organisation verarbeiten. Auch dieser Merkmalssatz enthält keine Auszeichnung.

Der Befund zur fehlenden Preisangabe verdient eine Zwischenbemerkung, weil er die Prämisse fast punktgenau adressiert. Der Preis wirkt hier als Inhalt, nicht als Angebotsauszeichnung. Es ist dieselbe Tatsache, aber in der Notation, die das Modell ohnehin liest.

Zhang, He und Yao (2026) messen nicht experimentell, sondern beobachtend, dafür in großer Breite. Sie werten 602 kontrollierte Anfragen über ChatGPT, Google AI Overview und Gemini sowie Perplexity aus, mit 21.143 gültigen Zitationen, 23.745 Merkmalsdatensätzen, 18.151 erfolgreich abgerufenen Seiten und 72 extrahierten Merkmalen. Der stärkste berichtete Zusammenhang mit dem Einfluss einer Seite auf die erzeugte Antwort ist ein modellgestützter Relevanzwert mit r gleich 0,4322, gefolgt von der Einbettungsähnlichkeit zwischen Antwort und zitierter Seite mit 0,3561. Ein Negativbefund wird von den Verfassern eigens hervorgehoben: Ein Frage-Antwort-Format allein verbessere die Aufnahme in die Antwort nicht. Für eine kausale Nachfolgestudie schlagen sie fünf Merkmalsfamilien vor, die einzeln variiert werden sollen, nämlich Struktur, Belegdichte, semantische Passung, Quellentransparenz und Seitenlänge. Eine Volltextsuche über die Arbeit ergibt für die Begriffe Schema, strukturierte Daten, JSON-LD und Auszeichnung keinen einzigen Treffer, auch nicht im Datenwörterbuch der 72 Merkmale. Die Verfasser schließen kausale Aussagen für ihre Datenlage ausdrücklich aus.

Martinez (2026) ordnet diese Einzelbefunde in einer kritischen Übersicht über fünfundvierzig Arbeiten aus dem Zeitraum November 2023 bis Juli 2026. Zwei Schlussfolgerungen sind für den vorliegenden Beitrag zentral. Erstens seien die viel zitierten Zugewinne der Gründungsarbeit innerhalb ihres Versuchsaufbaus gültig, aber daran gebunden, dass eine Quelle in einem festen Kontext bereits vorhanden sei; sie belegten weder organische Auffindbarkeit noch dauerhafte Zugriffswirkungen. Zweitens sei die Evidenzlage insgesamt schmal: Bereits abgerufener Inhalt könne seine Zitation kausal verändern, doch zeige keine der geprüften Techniken einen stabilen, längsschnittlichen, plattformübergreifenden kausalen Effekt auf organische Auffindbarkeit oder nachgelagertes Verhalten. Als reproduzierbarste Hebel benennt die Übersicht thematische Relevanz und Kontextposition.

Damit ist das dritte Glied nicht widerlegt, sondern unbelegt, und zwar auf eine spezifische Weise: Es ist nicht geprüft und gescheitert, sondern in der Experimentalliteratur gar nicht erst aufgestellt worden.

7. Die Position des Betreibers

Für Googles Oberflächen liegt eine Aussage vor, die keiner Interpretation bedarf. In der Anleitung zur Optimierung für generative Suchfunktionen findet sich ein Abschnitt, der ausdrücklich als Aufräumen mit Mythen überschrieben ist. Dort heißt es zu strukturier-

ten Daten, sie seien für die generative Suche nicht erforderlich, und es gebe keine besondere Schema.org-Auszeichnung, die man ergänzen müsse; es sei jedoch weiterhin sinnvoll, sie als Teil der allgemeinen Suchmaschinenoptimierung zu verwenden, weil sie zur Berechtigung für angereicherte Ergebnisse beitrage (Google, 2026). Zu maschinenlesbaren Zusatzdateien heißt es an gleicher Stelle, man müsse keine neuen maschinenlesbaren Dateien, KI-Textdateien, Auszeichnungen oder Markdown erstellen, um in der Google-Suche einschließlich ihrer generativen Funktionen zu erscheinen, da die Suche selbst sie nicht verwende. Die ältere Dokumentation zu den KI-Funktionen formuliert dasselbe knapper: Es gebe keine zusätzlichen Anforderungen und keine besonderen Optimierungen, um in KI-Übersichten oder im KI-Modus zu erscheinen (Google, 2025).

Diese Aussagen sind gewichtig und zugleich in ihrer Reichweite begrenzt. Sie sind Eigenauskunft eines Marktteilnehmers über die eigenen Systeme, keine unabhängige Messung. Sie gelten für Google und sagen nichts über ChatGPT, Perplexity, Copilot oder Claude. Und sie verwerfen die Auszeichnung nicht, sondern verschieben ihre Begründung: von der generativen Sichtbarkeit zurück auf die angereicherten Ergebnisse der klassischen Suche.

8. Gegenbefund und Verengung

Der stärkste Einwand gegen die hier vertretene Position stammt aus derselben Studie, die im vierten Abschnitt gegen die Lesbarkeit spricht. Tousseyn (2026) misst über 319 verfolgte Anfragen die Veränderung der Markenpräsenz nach der Schema-Einführung und findet über drei Monate für Googles KI-Übersichten eine Zunahme von 611 Prozent und für den KI-Modus von 42 Prozent, während ChatGPT um 71 Prozent, Copilot um 64 Prozent und Gemini um 35 Prozent zurückgingen und Perplexity unverändert blieb. Für die gesamte Website berichtet er über denselben Zeitraum eine Zunahme der Merkmale auf der klassischen Ergebnisseite um 377 Prozent und der Erscheinungen in KI-Übersichten um 1.500 Prozent.

Diese Zahlen sind kein Nebenbefund, sie sind der Gegenbefund, und sie verlangen eine Verengung der These. Auf Googles eigenen Oberflächen hat die Auszeichnung offenbar gewirkt. Die These, Auszeichnung sei für generative Sichtbarkeit ohne Belang, ist in dieser Allgemeinheit nicht haltbar.

Drei Präzisierungen halten den Befund und die These zugleich. Erstens hat der Studienautor selbst eine Kontrolle mitgeführt und berichtet, dass Wettbewerbermarken, die keine Auszeichnungsänderung vorgenommen hatten, im selben Zeitraum gleich große oder größere Bewegungen zeigten; er schließt daraus, dass ein erheblicher Teil des Anstiegs auf einen Plattformwechsel und nicht auf die Auszeichnung zurückgeht. Zwei-

tens deckt sich der verbleibende Effekt mit dem Mechanismus, den der Betreiber selbst angibt: Auszeichnung erzeugt Berechtigung für angereicherte Ergebnisse, angereicherte Ergebnisse gehören zum klassischen Index, und die KI-Übersichten setzen laut Google (2026) auf eben diesem Index auf. Der Wirkungspfad führt damit über die klassische Suche, nicht an ihr vorbei. Drittens trat der Effekt genau dort auf, wo dieser Pfad existiert, und nirgends sonst; auf den vier Systemen ohne Anbindung an Googles Index blieb er aus oder kehrte sich um.

Die These wird daher enger gefasst: Strukturierte Auszeichnung ist ein Hebel der klassischen Suche mit indirekter Wirkung auf jene generativen Oberflächen, die auf einem klassischen Index aufsetzen. Ein eigenständiger, plattformübergreifender Hebel generativer Sichtbarkeit ist sie nach gegenwärtiger Evidenz nicht.

Ein zweiter, schwächerer Gegenbefund verdient Erwähnung. Brinkmann et al. (2023) stellen fest, dass ein Vergleich der weit verbreiteten Eigenschaften mit jenen, die für die Aufnahme in Googles Schema.org-Anwendungen erforderlich sind, eine klare Korrelation zeige, was die Hypothese stütze, dass das Erscheinen in diesen Anwendungen eine wesentliche Motivation für die Auszeichnung sei. Wer Auszeichnung pflegt, tut dies also überwiegend dort, wo eine Belohnung sichtbar ist. Für die Interpretation von Korrelationen zwischen Auszeichnung und KI-Zitationen folgt daraus eine Warnung: Beide Größen können demselben Drittfaktor folgen, nämlich einer insgesamt gepflegten und in der klassischen Suche erfolgreichen Website.

9. Der Fall der maschinenlesbaren Zusatzdateien

Die Praxisliteratur empfiehlt neben der Auszeichnung zunehmend eigene maschinenlesbare Dateien, mit denen ein Unternehmen sein Angebot für Agenten verifizierbar veröffentlichen soll. Diese Empfehlung folgt derselben Denkfigur wie die Auszeichnungsprämisse und lässt sich an einem Beispiel prüfen, für das inzwischen Messdaten vorliegen.

Linehan und Guan (2026) untersuchten die Serverprotokolle von 137.210 Domains. Von diesen veröffentlichen 28 Prozent eine Indexdatei nach dem verbreiteten Vorschlag; 97 Prozent dieser Dateien erhielten im Beobachtungsmonat keinerlei Anfragen, weder von Bots noch von Menschen. Innerhalb der verbleibenden drei Prozent entfielen 1,1 Prozent der Anfragen auf jene Abrufbots, die live Antworten für Nutzeranfragen zusammenstellen, also auf genau die Systeme, deren Sichtbarkeit die Datei verbessern soll. Auf Anfragen an nicht vorhandene Dateien entfiel kein einziger Zugriff durch KI-Bots; die Verfasserinnen schließen daraus, dass kein System aktiv nach einer solchen Datei sucht. Sie weisen zugleich darauf hin, dass ihre Grundgesamtheit technisch versierter ist als das

Web insgesamt, dass die Verbreitungszahl daher eine Obergrenze darstellt, und dass ein Abruf noch keine Verwertung belegt.

Die Studie ist eine Anbieterstudie eines Werkzeugherstellers, der ein wirtschaftliches Interesse an der Debatte hat, und sie misst einen einzelnen Dateityp. Ihre Zahlen decken sich jedoch mit der Betreiberaussage, wonach die Suche solche Dateien nicht verwende (Google, 2026), und sie illustrieren ein Muster, das über den Einzelfall hinausweist: Ein Format setzt sich in der Optimierungspraxis durch, bevor geklärt ist, ob es gelesen wird. Linehan und Guan (2026) beschreiben dieses Muster in ihren Daten selbst, wenn sie berichten, dass Werkzeuge zur Prüfung und Katalogisierung der Datei zusammen 12 Prozent der Zugriffe erzeugen, also mehr als die Abrufbots und Assistenten zusammen.

10. Praktische Implikationen

Aus dem Befund folgen fünf Empfehlungen, die bewusst weniger versprechen als das geprüfte Praxisraster.

Erstens ist Auszeichnung zu pflegen, aber unter dem richtigen Titel. Sie gehört auf die technische Prüfliste der klassischen Suche und trägt dort belegt zur Berechtigung für angereicherte Ergebnisse bei (Google, 2026; Guha et al., 2016). Sie gehört nicht auf die Liste der Maßnahmen, von denen eine Verschiebung der Nennung in generativen Antworten erwartet wird.

Zweitens ist jede Aussage, die zählen soll, im sichtbaren Text zu führen. Wo eine Tatsache nur in der Auszeichnung steht, ist offen, ob sie überhaupt ankommt (Tousseyn, 2026), und der Betreiber verlangt ohnehin Übereinstimmung zwischen Auszeichnung und sichtbarem Text (Google, 2025). Der Preis, dessen Fehlen Vishwakarma et al. (2026) als Ausschlusskriterium messen, ist im Fließtext ein Preis; in der Auszeichnung ist er eine Zusatzangabe.

Drittens ist der Aufwand dorthin zu verlagern, wo gemessene Effekte liegen. Das sind thematische Passung und Kontextposition (Martinez, 2026), Belegdichte in Form von Quellenangaben, Zitaten und Zahlen (Aggarwal et al., 2024), Vollständigkeit und Aktualität (Vishwakarma et al., 2026) sowie semantische Nähe zwischen Frage und Seite (Zhang et al., 2026). Diese Liste ist unspektakulär und im Wesentlichen redaktionell.

Viertens ist bei jeder gemessenen Veränderung eine Wettbewerbskontrolle mitzuführen. Der instruktivste methodische Beitrag von Tousseyn (2026) ist nicht ein Ergebnis, sondern ein Verfahren: Bewegen sich Wettbewerber ohne die geprüfte Maßnahme gleich stark, liegt ein Plattformwechsel vor und keine Wirkung. Ohne diese Kontrolle ist jede Vorher-Nachher-Zahl in einem so bewegten Feld wertlos.

Fünftens ist Entitätsklarheit von ihrer Notation zu trennen. Dass ein Unternehmen über Quellen hinweg widerspruchsfrei und unterscheidbar beschrieben wird, bleibt eine sinnvolle Anforderung. Dass diese Beschreibung typisiert im Seitenkopf stehen müsse, um zu wirken, ist die Prämisse, die dieser Beitrag prüft und für unbelegt hält. Die Urheber des Vokabulars haben die dafür vorgesehene Verknüpfungseigenschaft selbst als kaum angenommen beschrieben (Guha et al., 2016).

11. Implikationen für Forschung

Die auffälligste Lücke ist zugleich die am einfachsten zu schließende. Zhang et al. (2026) skizzieren den passenden Aufbau bereits, wenn sie für eine kausale Nachfolgestudie fordern, an einer geschichteten Stichprobe von Seiten jeweils eine Merkmalsfamilie zu variieren und die Ergebnisse im selben Zeitfenster gegen dieselben Plattformen zu erheben. Eine solche Anlage, ergänzt um einen Arm, der ausschließlich die Auszeichnungsschicht verändert und den sichtbaren Text konstant hält, würde die hier geprüfte Prämisse in ihrem dritten Glied erstmals direkt testen. Dass dieser Arm bislang fehlt, ist eine Beobachtung des Verfassers über den Forschungsstand und wird als solche behauptet, nicht belegt.

Zwei weitere Fragen sind offen. Zum einen ist unklar, an welcher Stelle der Verarbeitungskette eingebettete Auszeichnung verloren geht, sofern sie verloren geht; die vorliegenden Hinweise stammen aus Verhaltensbeobachtung, nicht aus einer Vermessung der Abrufpfade. Zum anderen ist der Zusammenhang zwischen Auszeichnungsqualität und Zitationsverhalten unbearbeitet: Die Fehlerforschung endet vor der generativen Suche (Alrashed et al., 2021; Meusel & Paulheim, 2015), die Zitationsforschung beginnt nach ihr.

12. Limitationen

Der Beitrag ist konzeptionell angelegt und beansprucht keine eigene Messung. Er ordnet vorliegende Evidenz entlang einer selbst gebildeten Struktur und ist in seiner Aussagekraft durch diese Struktur begrenzt.

Die Belegbasis ist heterogen und in Teilen schwach. Martinez (2026) und Zhang et al. (2026) liegen als Preprints ohne Begutachtung vor; Zhang et al. schließen kausale Aussagen für ihre Daten ausdrücklich aus. Vishwakarma et al. (2026) ist begutachtet und auf einer Fachkonferenz erschienen, stammt aber vollständig von Mitarbeitenden eines Softwareanbieters, der die abgeleitete Prüfliste im eigenen Haus erprobt hat; ein Eigeninteresse ist damit nicht auszuschließen. Guha et al. (2016) ist begutachtet, aber Selbstauskunft der beteiligten Suchanbieter. Google (2025) und Google (2026) sind Betreiber-

dokumentation, also Eigenauskunft über die eigenen Systeme ohne unabhängige Prüfbarkeit.

Besonders zurückhaltend sind die beiden Anbieterstudien zu lesen. Tousseyn (2026) misst auf einer einzigen Domain eines Softwareunternehmens über drei Monate; der Autor selbst schränkt die Übertragbarkeit auf andere Branchen ein, und sein Fetch-Test prüft eine explizite Nutzeranweisung, nicht den regulären Abrufpfad. Linehan und Guan (2026) messen eine technisch überdurchschnittlich versierte Grundgesamtheit, weisen ihre Verbreitungszahl selbst als Obergrenze aus und halten fest, dass ein Abruf noch keine Verwertung belegt. Beide Studien stützen die hier vertretene Position, was ihre Gewichtung eher senkt als hebt; keine zentrale Behauptung dieses Beitrags ruht auf ihnen allein.

Ein methodisches Kernproblem betrifft den Argumenttyp. Die Feststellung, dass Auszeichnung in den Merkmalssätzen von Aggarwal et al. (2024), Vishwakarma et al. (2026) und Zhang et al. (2026) nicht vorkommt, ist ein Schluss aus Abwesenheit. Er belegt, dass die Wirkung nicht geprüft wurde, nicht, dass sie nicht besteht. Der Beitrag behauptet daher Unbelegtheit, nicht Unwirksamkeit, und diese Unterscheidung ist tragend.

Die zeitliche Reichweite ist eng. Die Beobachtungsfenster der aktuellen Arbeiten liegen zwischen Dezember 2025 und Juli 2026. Betreiberverhalten, Abrufpfade und Schnittstellen verändern sich in diesem Feld schnell; ein Wechsel der Verarbeitungspipeline eines einzigen großen Anbieters könnte den vierten Abschnitt hinfällig machen. Der Beitrag ist als Stand einer Debatte zu lesen, nicht als abschließendes Urteil.

Schließlich bleibt der Beitrag auf eine deutschsprachige, konzeptionelle Argumentationsebene beschränkt und integriert keine systematische Vollerhebung der internationalen Literatur; für eine Ausarbeitung in Richtung Zeitschriftenbeitrag wäre eine formalisierte Reviewmethodik erforderlich.

13. Fazit

Die Auszeichnungsprämisse hält der Prüfung in ihrer verbreiteten Fassung nicht stand, aber sie zerbricht auch nicht. Sie verengt sich.

Ihr erstes Glied, die Lesbarkeit, ist plattformabhängig und mindestens für mehrere große Antwortsysteme fraglich (Tousseyn, 2026). Ihr zweites Glied, die Notwendigkeit, widerspricht der Architektur retrieval-gestützter Systeme, die auf Textrepräsentationen arbeiten (Lewis et al., 2020), und dem dokumentierten Entstehungszweck des Vokabulars, das für schwächere Parser gebaut wurde (Guha et al., 2016). Ihr drittes Glied, die Wirkung auf die Nennung, ist in der Experimentalliteratur nicht geprüft worden, weil Auszeichnung in keinem der geprüften Merkmalssätze auftaucht (Aggarwal et al., 2024;

Vishwakarma et al., 2026; Zhang et al., 2026), und die verfügbare Übersichtsarbeit findet für keine geprüfte Technik einen stabilen plattformübergreifenden kausalen Effekt (Martinez, 2026).

Was bleibt, ist ein realer, aber indirekter Pfad. Auszeichnung erzeugt Berechtigung in der klassischen Suche, die klassische Suche speist die generativen Oberflächen jener Anbieter, die beides betreiben, und dort ist eine Wirkung gemessen worden (Tousseyn, 2026). Dieser Pfad ist schmaler als der verkaufte, aber er ist belegt, und er ist der einzige, der es ist.

Für die Praxis folgt daraus keine Entwarnung, sondern eine Umschichtung. Der Aufwand, der heute in Auszeichnungsschichten, Entitätsregister und maschinenlesbare Zusatzdateien fließt, hat seine gemessenen Entsprechungen in thematischer Passung, Belegdichte, Vollständigkeit, Aktualität und der Präsenz in Quellen, denen die Systeme ohnehin folgen. Das ist die unbequemere Empfehlung, weil sie sich schlechter als Produkt fassen lässt. Sie hat den Vorteil, dass die Evidenz sie trägt.

Literaturverzeichnis

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (S. 5–16). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671900

Alrashed, T., Paparas, D., Benjelloun, O., Sheng, Y., & Noy, N. (2021). Dataset or not? A study on the veracity of semantic markup for dataset pages. In *The Semantic Web – ISWC 2021* (Lecture Notes in Computer Science Bd. 12922, S. 338–356). Springer. https://doi.org/10.1007/978-3-030-88361-4_20

Brinkmann, A., Primpeli, A., & Bizer, C. (2023). The Web Data Commons schema.org data set series. In *Companion Proceedings of the ACM Web Conference 2023* (S. 1–4). Association for Computing Machinery. https://doi.org/10.1145/3543873.3587331

Google. (2025). *AI features and your website* [Dokumentation]. Google Search Central. https://developers.google.com/search/docs/appearance/ai-features

Google. (2026). *Optimizing your website for generative AI features on Google Search* [Dokumentation]. Google Search Central. https://developers.google.com/search/docs/fundamentals/ai-optimization-guide

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: Evolution of structured data on the web. *Communications of the ACM, 59*(2), 44–51. https://doi.org/10.1145/2844544

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33* (S. 9459–9474). Curran Associates. https://doi.org/10.48550/arXiv.2005.11401

Linehan, L., & Guan, X. (2026, 15. Juni). *We analyzed 137K sites: 97% of llms.txt files never get read* [Anbieterstudie]. Ahrefs. https://ahrefs.com/blog/llmstxt-study/

Martinez, O. (2026). *Optimizing visibility in generative engines: A critical survey of generative engine optimization (2023–2026)* (arXiv:2607.14035). arXiv. https://doi.org/10.48550/arXiv.2607.14035

Meusel, R., & Paulheim, H. (2015). Heuristics for fixing common errors in deployed schema.org microdata. In *The Semantic Web: Latest advances and new domains. ESWC 2015* (Lecture Notes in Computer Science Bd. 9088, S. 152–168). Springer. https://doi.org/10.1007/978-3-319-18818-8_10

Tousseyn, R. (2026, 23. März). *GEO experiment: Does schema markup really impact AI search?* [Anbieterstudie]. OtterlyAI. https://otterly.ai/blog/schema-markup-real-impact-ai-search/

Vishwakarma, R., Kumar, S., & Jamidar, R. (2026). What gets cited: Competitive GEO in AI answer engines. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery. https://doi.org/10.1145/3805712.3808445

Zhang, K., He, X., & Yao, J. (2026). *From citation selection to citation absorption: A measurement framework for generative engine optimization across AI search platforms* (arXiv:2604.25707v2). arXiv. https://doi.org/10.48550/arXiv.2604.25707

Hinweis zur Quellenlage. Bibliografische Angaben wurden einzeln gegen Crossref oder die Titelseite der Originalpublikation geprüft, nicht gegen Suchergebnis-Metadaten. Zu jeder Quelle liegt ein wörtliches Belegzitat aus der geöffneten Originalfassung vor. Belegklassen sind getrennt geführt und im Text kenntlich gemacht: begutachtet sind Aggarwal et al. (2024), Alrashed et al. (2021), Brinkmann et al. (2023), Guha et al. (2016), Lewis et al. (2020), Meusel und Paulheim (2015) sowie Vishwakarma et al. (2026); als Preprints ohne Peer Review liegen Martinez (2026) und Zhang et al. (2026) vor; Google (2025) und Google (2026) sind Betreiberdokumentation, also Eigenauskunft eines Marktteilnehmers; Linehan und Guan (2026) sowie Tousseyn (2026) sind Anbieterstudien und tragen keine zentrale Behauptung allein. Bei Guha et al. (2016) und Vishwakarma et al. (2026) besteht trotz Begutachtung ein Anbieterinteresse, das im Text und in den

Limitationen offengelegt ist. Der Begriff der Auszeichnungsprämisse in Abschnitt 3.1 ist eine eigene Begriffsprägung des Verfassers, ebenso die Feststellung der Forschungslücke in Abschnitt 11. Die Beobachtung, dass strukturierte Auszeichnung in den Merkmalssätzen der zitierten Experimentalarbeiten nicht vorkommt, ist ein Schluss aus Abwesenheit und als solcher gekennzeichnet; sie belegt fehlende Prüfung, nicht fehlende Wirkung. Wo eine Quelle für eine Aussage nicht ausreichte, wurde die Aussage gestrichen oder als eigene These gekennzeichnet; das Reparaturprotokoll ist Teil des Belegdossiers.