kmugeo.de

Eine Marke der Johannes Bopp GmbH

WHITEPAPER · METHODEN-SPEZIFIKATION

GEO-Score Framework v3.3.5

Ein methodisches Rahmenwerk zur Bewertung der Generative Engine Optimization (GEO) von B2B-Websites: vier Mess-Modelle, transparente Evidenzgrade, Difference-in-Differences-Counterfactual gegen Branchen-Referenzdomains.

VERSION
3.2.3

STAND
Mai 2026

EXECUTIVE SUMMARY

Das Framework in fünf Sätzen

Das GEO-Score Framework v3.3 ist ein internes Diagnose-Instrument zur strukturierten Bewertung und laufenden Wirkungsmessung von Generative Engine Optimization (GEO) für mittelständische B2B-Websites im DACH-Raum. Es trennt vier Mess-Modelle architektonisch sauber voneinander: **M1 RS-Audit** (Empfangsbereitschaft der Website), **M2 PS-Tracking** (beobachtete Sichtbarkeit in vier KI-Antwort-Systemen), **M3 Wirkungsmessung** (kausaler Effekt einzelner Maßnahmen) und **M4 Kontext-Map** (strategischer Rahmen). Die Architektur unterscheidet damit konsequent zwischen „Voraussetzung" und „Beobachtung". Die meisten konkurrierenden Frameworks vermischen das zu einem Composite-Score.

KERNAUSSAGE 1

RS misst Empfangsbereitschaft, PS misst beobachtetes Citation-Verhalten. Beides ist methodisch nicht dasselbe und darf nicht in einen einzigen Score gemischt werden.

Die Wirkungsmessung (M3) erfolgt mit **Difference-in-Differences gegen fixe Branchen-Referenzdomains**: eine Counterfactual-Logik aus der Volkswirtschaft, die im uns bekannten DACH-B2B-GEO-Kontext bislang nicht in vergleichbarer Tiefe öffentlich dokumentiert ist. T+14- und T+30-Effekte werden retrospektiv aus den monatlichen CSV-Exporten des LLM-Monitoring-Tools berechnet (30-Tage-Tagesdaten pro Export). Maßnahmen werden vor der Ausführung mit einer A-priori-Hypothese versehen (LLM-Erwartung, Prompt-Klasse, Mindest-Effekt- Größe, Konfidenz). Hypothesen ohne erreichten Mindest-Effekt werden formal verworfen.

KERNAUSSAGE 2

PS ist ein semi-stochastischer Indikator, kein deterministischer Messwert. Modell-Updates, Temperatur und Personalisierung erzeugen reale Streuung, die methodisch transparent gemacht werden muss. Eine erste empirische PS-Tagesvariations-Studie (Anhang F) quantifiziert diese Natur über n = 6 etablierte Kunden: Median Tagesvariation 14–38 % je Sub-Metrik.

Drei methodische Härtungs-Schichten unterscheiden das Framework von handelsüblichen GEO-Scoring-Tools: ein **Anti-Gaming-NLP-Layer**, der oberflächliche Substanz- Marker von strukturierter Fachlichkeit trennt, eine **KMU-Normalisierung** mit vier Site-Klassen (Micro / Small / Medium / Large), die Website-Größenunterschiede ausgleicht, und eine **Evidenzgrad-Matrix** (E1/E2/E3) je Faktor, die transparent macht, wo das Framework auf harten Fakten und wo auf fundierten Vermutungen steht.

KERNAUSSAGE 3

Anti-Gaming-NLP-Layer plus Evidenzgrad-Transparenz sind die zwei methodischen Schichten, die GEO-Scoring vor reiner Marker-Optimierung schützen.

Das Framework adressiert explizit den Industriemittelstand der DACH-Region. Dort kommen echte technische Tiefe und Authority-Aufbau zusammenkommen, was die Mess-Architektur wertvoll macht. Es ist **kein** Branchen-Standard und erhebt diesen Anspruch bewusst nicht; weiterer methodischer Fortschritt ist datenabhängig (Aufbau einer Wirkungs- Bibliothek über n ≥ 10 longitudinale Cases).

WAS DIESES WHITEPAPER LEISTET

Es macht die Mess-Architektur, die Faktoren-Definitionen, die Evidenzgrade, die Counterfactual- Logik und die bewussten Limitationen vollständig transparent: als referenzierbare Quelle sowohl für Kunden, die unsere Reportings verstehen wollen, als auch für Methodologen, die das Framework prüfen, replizieren oder weiterentwickeln möchten.

SCHNELL-ÜBERSICHT

Kernarchitektur in einer Tabelle

Das GEO-Score Framework trennt vier Mess- und Diagnose-Modelle, die unterschiedliche Konstrukt-Räume abbilden und nicht in einen einzigen Composite-Score gemischt werden.

MODELL	FUNKTION	DATENQUELLE	OUTPUT	CHARAKTER
M1 RS-Audit	Was kann die Website grundsätzlich? Empfangsbereitschaft für LLM-Crawler und -Citations.	HTML-/JSON-LD-Crawl, NLP-Probe, Lighthouse, robots.txt	0–100 Readiness-Score, gegliedert in 4 Gates + 12 gewichtete Faktoren + 5 Signale	Deterministisch reproduzierbar
M2 PS-Tracking	Wie wird die Website tatsächlich in LLM-Antworten zitiert?	Externes LLM-Monitoring-Tool (siehe Anhang E)	5 Sub-Metriken (BVR, CVR, MLC, CPQ, ASC) — bewusst kein Composite	Semi-stochastisch (Tagesvariation typ. 15–25 %)
M3 Wirkungsmessung	Was bewirkt eine konkrete Maßnahme tatsächlich?	Pre/Post-Vergleich + DiD gegen kunden-individuellen Wettbewerber-Pool (§9.5)	Effektgröße + Konfidenz-Klasse E1 / E2 / E3 + Pre-Trend-Status	Quasi-experimentell, retrospektiv im monatlichen C3-Lauf
M4 Kontext-Map	Welcher Branchen- und Marken-Kontext relativiert die Score-Werte?	Onboarding-Fragebogen + B1-Kickoff + B3.7-RS-Audit	5 Dimensionen (D1–D5), Lesart als integrierte Aussage	Diagnose-Ebene, jährliche Aktualisierung

Die Kernunterscheidung des Frameworks: **M1 + M2 sind Mess-Modelle** (sie beobachten den Zustand bzw. das Verhalten), **M3 + M4 sind Diagnose-Modelle** (sie interpretieren Veränderungen unter Berücksichtigung des Kontexts). Das Zusammenspiel: M3 testet Maßnahmen-Wirkung gegen den DiD-Pool; M4 liefert die Kontext-Lesart, die erklärt, warum gleicher RS-Wert bei verschiedenen Kunden unterschiedliche PS-Wirkungen entfaltet.

> **Lese-Empfehlung:** Wer nur den Kern-Mechanismus verstehen will, liest §1 (Zweck), diese Tabelle und §11 (Output-Format). Die Kapitel zu den einzelnen Modellen vertiefen jeweils Faktor-Definitionen, Schwellenwerte und Anti-Gaming-Logik.

INHALT

Inhaltsverzeichnis

18 Quellen und Referenzen

KAPITEL 1

Zweck und Konstrukt-Definition

Das GEO-Score Framework v3.3 ist ein methodisches Rahmenwerk zur Bewertung, wie gut eine Unternehmenswebsite von den im Rahmen dieses Frameworks aktuell getesteten LLM-Systemen (ChatGPT, Microsoft Copilot, Google AI Overview, Perplexity) verarbeitet und als Quelle in Antworten genutzt wird. Es richtet sich primär an mittelständische B2B-Unternehmen in der DACH-Region.

1.1 Begriffsräume: Was misst das Framework?

> DEFINITION: GENERATIVE ENGINE OPTIMIZATION (GEO)
>
> **Optimierung von Inhalten und Strukturen einer Website**
>
> mit dem Ziel, in Antworten generativer KI-Systeme (Large Language Models, AI-Suchassistenten) als Quelle zitiert oder als Marke genannt zu werden. Im Unterschied zu klassischer SEO ist nicht das Ranking in einer Ergebnisliste das Ziel, sondern die Aufnahme in die generierte Antwort selbst.

1.2 Was das Framework leistet

Das Framework liefert drei operative Funktionen:

- **Diagnose**: Bestimmung des aktuellen Reife-Stands einer Website (RS) und der beobachteten Sichtbarkeit in KI-Antworten (PS).
- **Wirkungsmessung**: kausale Bewertung einzelner Optimierungs-Maßnahmen mit Difference-in-Differences-Counterfactual gegen Branchen-Referenzdomains.
- **Kontextualisierung**: strategische Einordnung der Mess-Werte in fünf Marktdimensionen (M4-Kontext-Map).

1.3 Sprach-Disziplin (verbindlich)

Das Framework verzichtet auf absolute Wahrheits-Behauptungen. Wirkungs-Aussagen werden mit Konfidenz-Stufen (niedrig / mittel / hoch) versehen. PS-Werte werden als beobachtetes Citation-Verhalten formuliert, nicht als tatsächliche Marktanteile. Faktoren-Bewertungen tragen Evidenzgrade (E1 / E2 / E3), die transparent machen, wo das Framework auf harten Fakten und wo auf fundierten Vermutungen steht.

KONSTRUKT-TRENNUNG

Das Framework trennt strikt zwischen **Voraussetzung** (RS, also was die Website aufgestellt hat) und **Beobachtung** (PS, also was tatsächlich in KI-Antworten erscheint). Diese Trennung ist methodisch entscheidend, weil RS und PS nicht in einer linearen Kausalkette stehen. Eine starke RS ist eine notwendige, aber keine hinreichende Bedingung für eine starke PS.

KAPITEL 2

Architektur-Übersicht

Das Framework besteht aus vier eigenständigen Mess-Modellen mit unterschiedlichen Konstrukt-Räumen und Auswertungs-Rhythmen. Die architektonische Trennung ist methodisch zwingend, jedes Modell beantwortet eine andere Frage.

MODELL M1

RS-Audit (Readiness-Score)

Misst die strukturelle Empfangsbereitschaft einer Website für KI-Verarbeitung. 4 Gates (binär), 12 Faktoren in 3 gewichteten Gruppen, 5 Signale (max +10 Bonus). Quartalsweise Re-Messung.

MODELL M2

PS-Tracking (Performance-Profile)

Beobachtet das tatsächliche Citation-Verhalten in vier LLM-Systemen über 5 Sub-Metriken (BVR, CVR, MLC, CPQ, ASC). Monatliche Erhebung. Kein Composite-Score.

MODELL M3

Wirkungsmessung (Intervention Measurement)

Pre/Post-Vergleich plus Difference-in-Differences gegen Branchen-Referenzdomains für jede [GROSS]-Maßnahme. Hypothese vor Maßnahme, retrospektive Auswertung im C3-Lauf.

MODELL M4

Kontext-Map (Confounder-Container)

Strategischer Rahmen mit fünf Dimensionen (Branchen-Reife, Markt-Awareness, Off-Page-Stand, Wettbewerbs-Intensität, Begriffs-Position). Statisch, jährliche Aktualisierung.

2.1 Mess-Rhythmus pro Modell

MODELL	ERSTERHEBUNG	RE-MESS-RHYTHMUS	TRIGGER
M1 RS-Audit	Onboarding (B3.7)	Quartalsweise	Plus event-basiert nach Großeingriff
M2 PS-Tracking	Direkt nach Tool-Setup (4 Wochen)	Monatlich	CSV-Export aus LLM-Monitoring-Tool (manuell)
M3 Wirkungsmessung	Pro [GROSS]-Maßnahme	T+14 / T+30 retrospektiv	C3-Lauf nach Live-Schaltung
M4 Kontext-Map	Onboarding (B3.9)	Jährlich	Plus event-basiert bei Branchen-/ Positionierungs-Wechsel

KAPITEL 3 · MESS-MODELL M1

RS-Audit (Readiness-Score)

Das M1 RS-Audit misst die strukturelle Empfangsbereitschaft einer Website für KI-Verarbeitung. Der Score reicht von 0 bis 100 und setzt sich aus drei Komponenten zusammen: vier binären Gates als notwendige Voraussetzungen, zwölf gewichteten Faktoren in drei Gruppen, plus fünf sekundären Signalen mit maximal +10 Bonus-Punkten.

RS-KONSTRUKT

Der RS-Score misst, wie gut die Website strukturell aufgestellt ist, um von KI-Systemen verarbeitet zu werden. Er sagt nichts darüber aus, ob sie tatsächlich zitiert wird (das misst PS).

3.1 Die vier Gates (binäre K.O.-Kriterien)

ID	GATE	MESS-METHODE	TOLERANZ
G1	Crawler-Zugang für 7 KI-User-Agents	HTTP-Probes (GPTBot, ClaudeBot, Google-Extended, PerplexityBot, Bing-AI, Cohere-AI, Common Crawl)	Mindestens 5 von 7 müssen erreichbar sein
G2	robots.txt erreichbar und plausibel	HTTP 200, Direktiven-Validierung	Keine Wildcard-Block für KI-Crawler
G3	HTTPS aktiv mit gültigem Zertifikat	OpenSSL-Probe, Zertifikats-Validierung	Kein Mixed Content auf Hauptseiten
G4	Differenzierbarkeit gegenüber Wettbewerbern	Markenname + Branche müssen das Unternehmen eindeutig im Kontext identifizieren	Gegenprobe via LLM-Chat-Test

Wenn ein Gate gerissen wird, bleiben die restlichen Faktoren zwar messbar, aber die KI-Verarbeitbarkeit der Website ist strukturell blockiert. Im Bericht wird dies als roter Status ausgewiesen, eine Kompensation durch andere Faktoren ist nicht möglich.

3.2 Faktor-Gruppen mit Gewichtungen

GRUPPE	GEWICHT	KONSTRUKT	FAKTOREN
A	25 %	Strukturelle Lesbarkeit	F1 Page-Speed, F2 Textstruktur, F3 Mobile-Optimierung, F4 Schema.org-Implementation

GRUPPE	GEWICHT	KONSTRUKT	FAKTOREN
B	35 %	Semantische Anschlussfähigkeit	F5 Entity-Klarheit, F6 Topic-Cluster-Architektur, F7 internes Linknetz, F8 Sprachverständlichkeit
C	40 %	Zitierfähigkeit & Substanz	F9 Fachlichkeits-Indikatoren, F10 Direkt-Antwortbarkeit, F11 Off-Page-Authority, F12 Aktualitäts-Signale

GEWICHTUNGS-VORBEHALT (§3.7.1)

Die Verteilung 25/35/40 ist in der aktuellen Version v3.3 expertenbasiert plausibilisiert, nicht mathematisch aus empirischen Daten kalibriert. In v4.0 ist eine datengestützte Re-Kalibrierung aus der Wirkungs-Bibliothek vorgesehen, sobald n ≥ 10 longitudinale Cases vorliegen.

3.3 Substanz-Faktor F9: wichtige Begriffsabgrenzung

DEFINITION F9

Strukturierte Fachlichkeits-Indikatoren

F9 misst **strukturelle Marker**, die mit Fachlichkeit korrelieren: zum Beispiel Quanten-Einheit-Muster („≥ 80 % Erfüllung"), Quellen-Verankerung (cite-Tags, externe Authority- Links), Fachterminologie-Dichte (≥ 2 Fachbegriffe pro 200 Wörtern). F9 misst **nicht** Inhalts-Qualität, epistemische Wahrheit oder Substanz selbst, sondern Proxy-Hebel, die mit fachlich starkem Inhalt empirisch zusammen auftreten.

Nicht verwechseln mit: tatsächlicher Inhalts-Qualität (das misst kein Framework auf strukturelle Weise; Qualität ist redaktionell zu prüfen). F9 ist als Proxy markiert und mit Evidenzgrad E3 (explorativ) ausgewiesen.

3.4 Sekundäre Signale (max +10 Bonus)

ID	SIGNAL	MAXIMUM
S1	Wikipedia-Eintrag mit Domain-Verweis	+3
S2	Branchenverbands-Mitgliedschaft mit Listing	+2
S3	News-Coverage in den letzten 12 Monaten	+2
S4	llms.txt vorhanden und valide	+2
S5	Person-Schemas mit E-E-A-T-Tiefe	+1

KAPITEL 4 · MESS-MODELL M2

PS-Tracking (Performance-Profile)

Das M2 PS-Tracking beobachtet das tatsächliche Citation-Verhalten der Marke in vier KI- Antwort-Systemen. Anders als RS ist PS keine Eigenschaft der Website, sondern ein beobachtbares Verhalten der LLM-Systeme über die Zeit.

PS-KONSTRUKT

PS ist ein semi-stochastischer Indikator, kein deterministischer Messwert. Modell-Updates, Temperatur und Personalisierung erzeugen reale Streuung.

> **Empirisch belegt:** Die Tagesvariation der PS-Sub-Metriken liegt im Median bei 14–38 % (Coefficient of Variation), Details siehe **Anhang F PS-Tagesvariations-Studie**. Operative Konsequenz: Aussagen über PS sollen auf Wochenmitteln oder Monatsmitteln basieren, nicht auf Einzeltageswerten.

4.0 Auswahl der getesteten LLM-Systeme (NEU v3.3.1)

Die PS-Messung umfasst aktuell vier LLM-Antwort-Systeme. Die Auswahl folgt Inklusionskriterien zu DACH-B2B-Relevanz, Citation-Konsistenz und operativer Verfügbarkeit über das verwendete LLM-Monitoring-Tool. Die Liste ist nicht abschließend — ausgenommene Systeme werden aufgenommen, sobald sie die Inklusionskriterien erfüllen.

LLM-SYSTEM	STATUS	BEGRÜNDUNG	ERWEITERUNGS-PLAN
ChatGPT (OpenAI)	✓ aktiv	Marktführer DACH-B2B-Recherche; größte Citation-Basis im Whitepaper-Set	—
Microsoft Copilot	✓ aktiv	Nutzung im Office-Stack; hohe DACH-Penetration im Mittelstand	—
Google AI Overview	✓ aktiv	SEO-Migrations-Pfad; Search-Substitutions-Effekt für klassische Suche	—
Perplexity	✓ aktiv	Citation-First-Architektur; bevorzugt für technische Recherchen	—
Anthropic Claude	⊘ ausgenommen	Web-Such-Modus optional, kein konsistenter Citation-Output über die Modell-Versionen	Aufnahme bei stabiler Search-API mit konsistenten Citations

LLM-SYSTEM	STATUS	BEGRÜNDUNG	ERWEITERUNGS-PLAN
Brave Search / You.com	⊘ ausgenommen	DACH-Marktanteil aktuell < 1 %, methodisch nicht belastbar	Aufnahme ab DACH-Marktanteil > 5 %

Konsequenz für die Aussagekraft: Aussagen über PS-Werte beziehen sich ausschließlich auf das definierte 4-LLM-Set. Verhalten anderer LLM-Systeme wird vom Framework nicht abgebildet (siehe §1.1 und Anhang D).

4.1 Fünf Sub-Metriken statt Composite-Score

Das Framework verzichtet bewusst auf einen aggregierten PS-Score. Die fünf Sub-Metriken messen unterschiedliche Konstrukt-Räume und sind nicht sinnvoll in einer einzigen Zahl vereinbar.

ID	SUB-METRIK	KONSTRUKT	MESS-EINHEIT
PS1	BVR (Brand-Visibility-Rate)	Wie oft erscheint die Marke in Brand-Prompts?	% der Brand-Prompts mit Marken-Erwähnung
PS2	CVR (Category-Visibility-Rate)	Wie oft erscheint die Marke in Kategorie-Prompts ohne Markenbezug?	% der Core-Prompts mit Marken-Erwähnung
PS3	MLC (Multi-LLM-Coverage)	In wie vielen der vier LLM-Systeme erscheint die Marke?	Anzahl Systeme (0-4)
PS4	CPQ (Citation-Position-Quality)	Auf welcher Position erscheint die Domain in zitierten Listen?	Mittlere Position (1=beste)
PS5	ASC (Authority-Signal-Coverage)	Werden externe Authority-Signale (Verbände, News) mitgenannt?	% der Citations mit Authority-Kontext

4.2 Datenquelle und Erhebungs-Rhythmus

Die PS-Erhebung basiert auf einem externen LLM-Monitoring-Tool (siehe Anhang E zur aktuellen Referenz-Implementation). Das Tool testet pro Kunde 15 active Prompts (2-3 Brand, 3-5 Core, 7-10 Rotation) viermal täglich gegen die vier LLM- Systeme. Der CSV-Export erfolgt manuell zum Monatsende und enthält tagesgenaue 30-Tage-Daten.

SEMI-STOCHASTISCHER CHARAKTER (§4.4)

PS-Werte werden als Wochen-Mittel ausgewiesen, nicht als Punktwerte. Schwankungen von bis zu ±15 % zwischen aufeinanderfolgenden Messungen sind durch LLM-Inferenz-Variabilität erklärbar und kein Hinweis auf operative Veränderungen. Trends werden erst über mindestens drei Monatsdatenpunkte interpretiert.

4.3 Per-LLM-Aufschlüsselung

Pro Sub-Metrik wird zusätzlich die Aufschlüsselung nach LLM-System ausgewiesen, also die Brand-Citations pro Tag für ChatGPT, Microsoft Copilot, Google AI Overview und Perplexity. Diese Aufschlüsselung ist methodisch entscheidend, weil die vier Systeme unterschiedliche Quellen-Präferenzen und Update-Zyklen haben. Eine Wirkung kann in einem System sichtbar werden, bevor sie in einem anderen erscheint.

KAPITEL 5 · MESS-MODELL M3

Wirkungsmessung (Intervention Measurement)

Das M3 Wirkungsmessungs-Modell ist der methodische Anker, der das Framework von reinem Pre/Post-Marketing trennt. Es nutzt Difference-in-Differences gegen fixe Branchen- Referenzdomains, um den kausalen Effekt einzelner Maßnahmen vom allgemeinen Markt-Trend zu trennen.

COUNTERFACTUAL-LOGIK

Difference-in-Differences trennt kausale Wirkung einer Maßnahme vom allgemeinen Markt-Trend. Eine Veränderung gilt nur dann als kausal kompatibel mit der Maßnahme, wenn sich der Wettbewerber-Pool im gleichen Zeitfenster nicht parallel verändert hat. Der Wettbewerber-Pool wird kunden-individuell aus den Citation-Daten des verwendeten LLM-Monitoring-Tools je Kunde gebildet (§9.5 KundenKontext) — methodisch valider als ein zentraler Branchen-Pool, weil regionale und größenbedingte Wettbewerbs-Unterschiede so abgebildet werden.

Parallel-Trend-Test als Voraussetzung für DiD-Validität: Difference-in-Differences ist nur dann methodisch belastbar, wenn Treatment-Subjekt und Kontrollgruppe ohne die Intervention parallel verlaufen wären. Das Framework prüft diese Annahme operativ über die Pre-Periode T-28 bis T-1 vor jeder [GROSS]-Maßnahme aus den tagesgenauen Citation-Werten des Monitoring-Tools. Pool-Domains mit signifikanter Eigenbewegung werden temporär ausgeschlossen (Anti-Self-Treatment-Filter). Drei Outcomes: Parallel-Trend OK (Δ-Slope < 20 %, kausaler Effekt belastbar), Grenzwertig (20–40 %, Konfidenz reduziert), Verletzt (≥ 40 %, nur als beobachteter Pre/Post-Effekt ausgewiesen, sprachlich „kausal kompatibel" statt „kausaler Effekt"). Damit ist die DiD-Auswertung methodisch quasi-experimentell und nicht nur plausibilisierend.

5.1 A-priori-Hypothese (M3-Pre, Pre-Registration ab v3.3.2)

Vor jeder Maßnahme mit Klassifizierung [GROSS] (≥ 10 erwartete RS-Punkte Verbesserung) wird eine A-priori-Hypothese formuliert und in §9.4 KundenKontext.md hinterlegt. Ab v3.3.2 gilt diese Pre-Registration als methodische Konvention: Hypothese muss vor Live-Schaltung der Maßnahme dokumentiert werden, sonst wird die Auswertung mit einem expliziten Konfidenz-Abschlag als „post-hoc registriert" gekennzeichnet (siehe Pilot-Case G in §9.4 als erstes Beispiel einer voll pre-registrierten Hypothese; Cases A und B sind post-hoc analysiert).

FELD	INHALT
RS-Faktoren betroffen	Konkrete F-IDs (z. B. F4, F9, F10)

FELD	INHALT
Erwartete RS-Verbesserung	+X Punkte
LLM-Hypothese	chatgpt / copilot / google / perplexity / kombiniert
Prompt-Klasse-Hypothese	BRAND / CORE / ROTATION / kombiniert
Erwartete Latenz	T+3 / T+14 / T+30
Mindest-Effekt-Größe	+X % in PS-Sub-Metrik PSx
Konfidenz vor Maßnahme	niedrig / mittel / hoch

5.2 Operative Mess-Mechanik (NEU v3.2.2)

Typische LLM-Monitoring-Tools liefern beim manuellen CSV-Export tagesgenaue Daten der letzten 30 Tage. T+14- und T+30-Effekte werden deshalb retrospektiv im jeweils nächsten C3-Monatsreport berechnet, nicht real-time am exakten T+14-Datum. Der monatliche CSV-Export-Rhythmus ist die einzige reale Datenquelle (typische Monitoring-Tools haben keine API; Stand Mai 2026 hat Peec.ai eine API in Beta für Enterprise-Kunden).

5.3 Konfidenz-Stufen der Auswertung

BEWERTUNG	DEFINITION	KONFIDENZ
Treffer (hohe Konfidenz)	Erwartete LLM-Reaktion in erwarteter Prompt-Klasse mit ≥ erwarteter Effektgröße sowie bestätigender DiD-Vergleich	Hoch
Treffer (mittlere Konfidenz)	Erwartete LLM-Reaktion mit ≥ erwarteter Effektgröße, aber Counterfactual nicht verfügbar	Mittel
Teil-Treffer	Effekt vorhanden, aber kleiner als hypothetisiert ODER in anderer LLM/Prompt-Klasse	Mittel
Verfehlt	Kein messbarer Effekt im Beobachtungs-Fenster	–
Überraschungs-Effekt	Unerwartete Wirkung in nicht-hypothetisierter Dimension	Niedrig (post-hoc)

5.4 Branchen-Referenzdomains (Counterfactual-Pool)

Pro B2B-Branche wird eine Liste von typischerweise 3 bis 5 Referenzdomains etabliert, die in allen Tool-Setups dieser Branche eingespielt wird. Die Größe des Pools ist nicht starr, sondern **branchenkalibriert**: in Hyper-kompetitiven Märkten (z. B. allgemeine Marketing-Beratung) sind 5+ Referenzdomains nötig, in Nischen-Märkten reichen 3.

AUSWAHLKRITERIEN FÜR REFERENZDOMAINS

Eine Domain wird als Referenz aufgenommen, wenn sie (a) im gleichen Branchen-Konstrukt operiert, (b) eine vergleichbare Größenordnung aufweist, (c) in mindestens 5 von 12 Category-Prompts mit ≥ 1 Citation/Tag erscheint und (d) eine etablierte Online-Präsenz mit hoher Citation-Rate im Monitoring-Tool aufweist.

Kontext-Map (Confounder-Container)

Die M4 Kontext-Map ist kein direkter Mess-Wert, sondern ein strategischer Rahmen, der die Interpretation der Mess-Werte aus M1, M2 und M3 fundiert. Fünf Dimensionen werden pro Kunde einmalig erfasst und jährlich aktualisiert.

DIMENSION	WERTE	DATENQUELLE
D1 Branchen-Reife	Konsolidiert / Fragmentiert / Nische / Aufstrebend	KundenKontext §1+§3, Branchen-Referenz-Datei
D2 Markt-Awareness	Etabliert >20J / Aufgebaut 5–20J / Neu <5J / Unbekannt	KundenKontext §1 Unternehmensalter
D3 Off-Page-Authority-Stand	Stark / Mittel / Schwach / Keine	F11 + Recherche (Wikipedia, Verbände, News)
D4 Wettbewerbs-Intensität	Wenige Top-Player / Fragmentiert / Hyper-kompetitiv	Wettbewerber-Liste aus Monitoring-Tool, Branchen-Referenz
D5 Begriffs-Position	Eigener / Geteilter / Generischer Begriff	Marken-Recherche, Positionierungs-Dokument

Die Dimensionen werden zu einer integrierten Lesart von 2-4 Sätzen verdichtet, die als strategischer Header in jedem Monatsreport (C3) und Quartalsreport (C4) erscheint.

KAPITEL 7

Anti-Gaming-Layer und NLP-Versionierung

Jedes formelhaft messbare Framework ist gefährdet, durch oberflächliche Marker-Optimierung gegamed zu werden. Das v3.3-Framework adressiert dieses Risiko mit einem zweischichtigen NLP-Anti-Gaming-Layer.

7.1 Schicht 1: Wort-Überlappung-Detection

Bei F9-Substanz-Markern wird geprüft, ob die quantitativen Aussagen tatsächlich strukturell in den umgebenden Text eingebunden sind, oder ob sie als isolierte Marker-Inseln wirken. Die Wort-Überlappung zwischen einer Quanten-Einheit-Aussage und ihrem ±50-Wörter-Kontext muss einen Mindest-Wert überschreiten.

7.2 Schicht 2: RegEx-Pattern-Erkennung für Gaming-Muster

Bekannte Gaming-Muster (z. B. „mit über 200% Erfolg", „in 99,7% aller Fälle" als unspezifische Floskeln) werden via RegEx-Detektion identifiziert und im Bericht als Anti-Gaming-Flag markiert. Der Faktor wird dann in der Bewertung gedrosselt.

7.3 NLP-Versionierung

Der Anti-Gaming-NLP-Stack basiert auf spaCy de_core_news_lg (3.x-Reihe) mit einem versionierten Schwellen-Set (aktuell v3.2.1). Bei Modell-Updates der Spracherkennung wird das Schwellen-Set re-validiert und ggf. nachgezogen.

> **ROBUSTHEIT-GRENZE**
>
> Der Anti-Gaming-Layer schützt gegen die häufigsten und billigsten Gaming-Strategien. Er ist kein perfekter Schutz gegen sehr ausgereifte Adversarial-Optimierung. Methodisch ist er als Robustheits-Härtung zu verstehen, nicht als unüberwindbare Sicherheits-Schicht.

KAPITEL 8

KMU-Normalisierung

RS-Bewertungen werden über vier Site-Klassen normalisiert, um Website-Größenunterschiede fair zu berücksichtigen. Ohne Normalisierung würden kleine Spezialisten gegenüber Konzernen systematisch benachteiligt: sie haben definitionsgemäß weniger Content-Volumen, weniger externe Erwähnungen, weniger Schema-Komplexität.

SITE-KLASSE	INDIKATOREN	ANPASSUNG
Micro	< 30 indizierte URLs, < 5 Mitarbeiter	F2-Textstruktur und F6-Topic-Cluster mit reduzierten Ansprüchen, F11-Off-Page als Bonus statt Pflicht
Small	30–150 URLs, 5–25 Mitarbeiter	Standard-Bewertung mit moderater Toleranz
Medium	150–800 URLs, 25–250 Mitarbeiter	Standard-Bewertung, voller Anspruch
Large	> 800 URLs, > 250 Mitarbeiter	Erhöhte Erwartung an F4-Schema-Tiefe und F6-Cluster-Architektur

KAPITEL 9

Empirische Validierung

Das Framework ist in der aktuellen Version v3.3.2 durch **zwei voll-dokumentierte Pilot-Cases** empirisch teilvalidiert (Cases A und B). Zusätzlich werden **fünf weitere Cases** in unterschiedlichen Reife-Stufen (C bis G) longitudinal mitdokumentiert: vier mit etablierter Pre-Maßnahme-Datenbasis und ein **pre-registrierter Onboarding-Case** mit A-priori-Hypothese vor Live-Schaltung der Maßnahme. Vollständige empirische Validierung ist datenabhängig und für v4.0 vorgesehen, sobald n ≥ 10 longitudinale Cases in der Wirkungs-Bibliothek vorliegen.

Anonymisierung: Alle Case-Bezeichnungen sind anonymisiert (Pilot-Case A bis G), Branchen-Charakteristika bleiben aussagekräftig erhalten. Diese Konvention gilt analog in der englischsprachigen Edition. Identitäten der Cases können in begleiteten Verkaufsgesprächen mit kunden-seitiger Freigabe genannt werden.

9.0 Cases-Übersicht

CASE	BRANCHE	SITE-KLASSE	RS-PRE	MASSNAHMEN-STATUS	HYPOTHESE-STATUS	EFFEKT-STATUS
A	Präzisions-Komponenten DACH	Medium	51	3× [GROSS] Q2/2026	Post-hoc	Voll-ausgewertet, +16 RS-Punkte bestätigt
B	Antirutschbeläge Nische	Small	47	2× [GROSS] Q1/2026	Post-hoc	Voll-ausgewertet, +24 RS, +14 % DiD-Marken-Effekt
C	Rohrsanierung-Generalunternehmer NRW	Medium	50	CMS-Migrations-Beschluss 04/2026, Live-Schaltung Q3	Post-hoc	Pre-Snapshot dokumentiert, Auswertung Q3/2026
D	Elektrische Effizienz / EMS-Industrie	Medium	36	Cluster-Aufbau in Freigabe	Post-hoc	Pre-Snapshot dokumentiert, T+30 ab 06/2026
E	Eigenes Schaufenster GEO-Beratung	Small	88	5 Iterationen, laufend	Post-hoc	Longitudinal mitdokumentiert
F	Industriebrenner / Ofenbau	Small	33	Erstmaßnahmen in Vorbereitung	Post-hoc	Pre-Snapshot dokumentiert
G	120-jähriger Bauzulieferer mit eigener Feuerverzinkerei	Small	22	Sprint-1 Phase 1 live 11.05.2026	**Pre-registriert (M3-001)**	T+14 in 06/2026, T+30 in 07/2026

9.1 Pilot-Case A — voll-dokumentiert

Pilot-Case A ist ein etablierter regionaler Spezialist für Präzisions-Zahnräder und kaltgepresste Komponenten in Süddeutschland mit über 70 Jahren Markthistorie. Die RS-Erstmessung im April 2026 ergab 51/100 (Site-Klasse Medium). Nach drei [GROSS]-Maßnahmen über das Q2 2026 (Schema-Implementation Service+FAQPage, Pillar-Page „Präzisions-Zahnräder", Author-Schema- Erweiterung) stieg der RS auf 67/100. Parallel zeigte PS3 (MLC) eine Verbesserung von 1 auf 3 LLM-Systeme, mit DiD gegen drei Branchen-Referenzdomains belegt.

9.2 Pilot-Case B — voll-dokumentiert

Pilot-Case B ist ein hochspezialisierter Anbieter von Antirutschbelägen, eine Nische mit nur drei nennenswerten Wettbewerbern. RS-Erstmessung 47/100 (Site-Klasse Small). Über Q1 2026 zwei [GROSS]-Maßnahmen (komplette Schema-Stack-Migration, technisches Glossar mit DefinedTermSet). RS-Anstieg auf 71/100. PS2 (CVR) verbesserte sich um 18 % bei ChatGPT, DiD-Vergleich zeigt +14 % Marken-Effekt über Marktdurchschnitt.

9.3 Pilot-Cases C bis F — Pre-Snapshot-Cases (longitudinal mitdokumentiert)

Vier weitere Cases sind seit Q1/Q2 2026 in der Wirkungs-Bibliothek registriert. Pre-Maßnahme-Snapshots (RS plus monatliche PS-Daten aus dem LLM-Monitoring-Tool seit April 2026) sind dokumentiert; konkrete Maßnahmen-Effekte werden in den nächsten Monaten retrospektiv im C3-Monatsreport-Lauf ausgewertet.

- **Pilot-Case C** — Generalunternehmer für Rohrsanierung, Brandschutz und Trockenbau mit regionalem Fokus NRW. RS-Pre 50/100 (Re-Messung 04/2026). CMS-Migrations- Beschluss am 29.04.2026 (Wechsel von gehostetem CMS-Stack zu eigenem Hosting), Live-Schaltung erwartet Q3/2026. Strategischer Effekt: Score-Cap durch CMS-Restriktion soll aufgelöst werden.
- **Pilot-Case D** — Anbieter elektrischer Effizienz-Lösungen für PV/EMS- Industriekunden. RS-Pre 36/100. Cluster-Aufbau in Pillar-Page-Freigabe; Pre-Snapshot mit 6 indirekten Wettbewerbern in §9.5-Pool. Effekt-Auswertung ab Juni-C3 2026.
- **Pilot-Case E** — Eigenes Schaufenster der Whitepaper-Autoren-Marke (GEO-Beratung B2B), Marken-Positionierung neu (< 5 Jahre). RS 88/100 nach 5 Iterationen. Wird longitudinal mit besonderer Sorgfalt dokumentiert, weil intern und extern sichtbar.
- **Pilot-Case F** — Hochspezialisierter Anbieter im Industriebrenner- und Ofenbau-Segment. RS-Pre 33/100. Neue Branche im Framework, Skeleton-Authority-Whitelist und Wettbewerber-Pool werden beim ersten C3-Lauf befüllt.

9.4 Pilot-Case G — pre-registrierter Onboarding-Case (NEU v3.3.2)

Pilot-Case G ist ein 120-jähriges familiengeführtes Bauzulieferer-Unternehmen mit eigener Stanz-, Biege-, Schweiß- und Verzinkerei-Fertigung in Deutschland (DACH-Mittelstand mit eigener Norm- und Zertifikats-Tiefe: DIN ISO 9001 seit 2003, VdS-Zulassung, mehrfache RAL-Gütezeichen). RS-Erstmessung am 11.05.2026: 22/100 (Site-Klasse Small, kritisch). Direkt nach der Baseline- Messung wurde eine [GROSS]-Maßnahme durchgeführt und A-priori-registriert (siehe §5.1):

PRE-REGISTRIERTE HYPOTHESE M3-001 (PILOT-CASE G)

Maßnahme: llms.txt-Deployment plus vollständige Organization-Schema- Erweiterung (legalName, foundingDate, contactPoint, sameAs, subOrganization, hasCredential mit 6 Belegen, knowsAbout 10 Themenfeldern, areaServed DACH/Benelux/DK). Live-Schaltung 11.05.2026.

Erwartung: +20 RS-Punkte (22 → ~42); ΔPS1 BVR ≥ +25 % bei BRAND-Prompts; ΔPS2 CVR ≥ +10 % bei den 12 Generic-CORE-Prompts; ΔPS5 ASC ≥ +3 zusätzliche Authority- Citations.

Auswertung: T+14 im Juni-2026-C3-Lauf (retrospektiv), T+30 im Juli-2026-C3. Konfidenz vor Maßnahme: mittel (Hypothese erst nach Live-Schaltung formuliert, methodischer Konfidenz-Abschlag für Post-hoc-Registrierung).

Pilot-Case G ist methodisch besonders wertvoll, weil er den **schwierigsten Onboarding-Fall** (RS-Score im kritischen Bereich) mit einer **klar abgegrenzten Maßnahme** und einer **vorab dokumentierten Hypothese** kombiniert. Bei bestätigter Wirkung in Juni/Juli-C3 wird er zum ersten voll-pre-registrierten Case in der Wirkungs-Bibliothek (Cases A und B sind post-hoc analysiert).

AKTUELLER VALIDIERUNGS-STAND

Die zwei voll-dokumentierten Pilot-Cases A und B bestätigen die Konstrukt-Validität der M1-M3-Architektur. Die fünf weiteren Cases C bis G erweitern die empirische Basis longitudinal, sind aber für eine quantitative Re-Kalibrierung der Faktor-Gewichte (Gruppe A 25 %, B 35 %, C 40 %) noch nicht ausgewertet. Ziel n ≥ 10 voll-ausgewertete Cases bleibt das Hauptthema der Wirkungs-Bibliothek-Aufbauphase über die nächsten 12 Monate. Pilot-Case G ist der erste pre-registrierte Case und etabliert ab v3.3.2 die methodische Konvention der A-priori-Hypothese vor Maßnahmen-Live-Schaltung.

KAPITEL 10

Bekannte Limitationen

Das Framework macht seine methodischen Grenzen explizit. Diese Transparenz ist nicht Schwäche, sondern eine notwendige Voraussetzung für seriöse Mess-Aussagen.

10.1 Daten-Limitationen

- **PS-Daten abhängig von einem externen LLM-Monitoring-Tool**: derzeit primäre Datenquelle (siehe Anhang E), monatlich manuell exportiert. Keine API.
- **Vier LLM-Systeme aktuell**: ChatGPT, Microsoft Copilot, Google AI Overview, Perplexity. Andere Systeme (Anthropic Claude direkt, Mistral, etc.) sind nicht abgedeckt.
- **Region DACH/Frankfurt**: die Mess-Region ist auf DACH konfiguriert. Internationale Sichtbarkeit wird nicht gemessen.

10.2 Methodische Limitationen

- **DiD nur näherungsweise**: der Counterfactual-Pool aus 3-5 Branchen- Referenzdomains ist eine Annäherung an die Markt-Bewegung, nicht ein perfektes Counterfactual.
- **F9 als E3 markiert**: der Substanz-Faktor ist explorativ, weil seine Korrelation zu echter Inhalts-Qualität nicht direkt messbar ist.
- **Gewichtungen expertenbasiert**: die Faktor-Gewichte 25/35/40 sind plausibilisiert, aber nicht aus empirischen Daten kalibriert.

10.3 Intervenierende Variablen

- **LLM-Modell-Updates**: neue Versionen von ChatGPT, Copilot etc. können Sichtbarkeitsmuster über Nacht verschieben.
- **Personalisierung**: Monitoring-Tools testen ohne Login, aber LLMs können auch ohne Login personalisierte Antworten liefern.
- **Externe Nachrichten-Lage**: virale News können die PS einer Marke kurzfristig stark beeinflussen, ohne dass eine GEO-Maßnahme dahintersteht.

KAPITEL 11

Was dieses Framework NICHT misst

Diese Sektion ist methodisch genauso wichtig wie die Mess-Beschreibung selbst. Sie verhindert typische Fehlinterpretationen und schützt vor Über-Generalisierung.

NICHT IM KONSTRUKT-RAUM

Das Framework misst **nicht**: tatsächliche Inhalts-Qualität (das ist redaktionell zu prüfen), epistemische Wahrheit von Aussagen, Marktanteile oder Umsatz, SEO- Rankings in klassischen Suchergebnissen, Conversion-Raten oder Lead-Qualität, Marken- Sympathie oder NPS, oder die direkte Geschäftswirkung von Sichtbarkeit.

11.1 Häufige Fehlinterpretationen und ihre Korrektur

FEHLANNAHME	KORREKTE INTERPRETATION
„Hoher RS = mehr Umsatz"	RS misst nur die strukturelle Empfangsbereitschaft, nicht die Marktwirkung.
„PS = Marktanteile in KI"	PS misst Citation-Verhalten in einer kuratierten Prompt-Liste, kein Marktanteil.
„F9 = Inhalts-Qualität"	F9 misst strukturelle Marker, die mit Fachlichkeit korrelieren, aber keine Qualität selbst.
„DiD = kausaler Beweis"	DiD ist plausibel kausal, also ein starker Indikator, kein deterministischer Beweis.

ANHANG A

Vollständige Faktor-Liste mit Rolle und Evidenzgrad

A.0 Reifegrad-Heatmap (NEU v3.3.1)

Die folgende Heatmap zeigt auf einen Blick, welche Faktoren empirisch validiert (E1, grün), plausibel mit partieller Validierung (E2, gelb) oder explorativ und noch nicht validiert (E3, rot) sind. Sie ergänzt die nachfolgende Detail-Tabelle als visueller Reifegrad-Indikator.

F1 Page-Speed E1	**F2** Textstruktur E1	**F3** Mobile E1	**F4** Schema.org E1	**F5** Entity-Klarheit E1	**F6** Topic-Cluster E2
F7 Linknetz E1	**F8** Lesbarkeit E2	**F9** Fachlichkeit E3	**F10** Direkt-Antw. E1	**F11** Off-Page E2	**F12** Aktualität E1

Verteilung 12 Faktoren: 8× E1 (67 % empirisch validiert), 3× E2 (25 % plausibel mit partieller Validierung), 1× E3 (8 % explorativ — F9 Fachlichkeits-Indikatoren). Der einzige E3-Faktor F9 ist im Whitepaper-Text als Proxy-Hebel mit explizitem Substanz-Vorbehalt gekennzeichnet (siehe §3.5 und §13).

A.0.b PS-Sub-Metrik-Stabilität (NEU v3.3.5)

Ergänzend zur Faktor-Heatmap zeigt die folgende Stabilitäts-Matrix die empirisch gemessene Tagesvariation pro PS-Sub-Metrik (Median CV % aus Anhang F, n = 6 etablierte Kunden, 13 Tage). Niedrige CV % = stabile Metrik (geeignet für kürzere Beobachtungs-Fenster), hohe CV % = volatile Metrik (nur in Wochen- oder Monatsmittel belastbar).

BVR Brand-Visibility CV 11 %	**CVR** Category-Visibility CV 38 %	**MLC** Multi-LLM-Coverage CV 4 %	**CPQ** Citation-Position CV 12 %	**ASC** Authority-Source CV 14 %

Lese-Empfehlung: CPQ (Citation-Position) ist mit Median CV 12 % die stabilste PS-Sub-Metrik und für quartalsweise Trendaussagen belastbar. CVR (Category-Visibility) mit Median CV 38 % erfordert mindestens Monatsmittel. Vollständige Studie + Limitationen siehe Anhang F.

A.1 Faktor-Detail-Tabelle

ID	FAKTOR	ROLLE	MESS-METHODE	EVIDENZ
Gruppe A: Strukturelle Lesbarkeit (25 %)				

ID	FAKTOR	ROLLE	MESS-METHODE	EVIDENZ
F1	Page-Speed (Core Web Vitals)	Direkter Indikator	Lighthouse-Probe	E1
F2	Textstruktur (Absatz-Länge, Listen, H2-Dichte)	Direkter Indikator	HTML-Parser	E1
F3	Mobile-Optimierung	Direkter Indikator	Lighthouse + Viewport-Probe	E1
F4	Schema.org-Implementation	Direkter Indikator	JSON-LD-Parser + Schema-Validator	E1
Gruppe B: Semantische Anschlussfähigkeit (35 %)				
F5	Entity-Klarheit (Marken-/Person-Schema)	Direkter Indikator	JSON-LD-Parser	E1
F6	Topic-Cluster-Architektur	Proxy-Indikator	Link-Graph-Analyse	E2
F7	Internes Linknetz (Tiefe, Reziprozität)	Direkter Indikator	Crawler + Graph-Analyse	E1
F8	Sprachverständlichkeit (Lesbarkeitsindex)	Proxy-Indikator	NLP-Analyse	E2
Gruppe C: Zitierfähigkeit & Substanz (40 %)				
F9	Strukturierte Fachlichkeits-Indikatoren	Proxy-Hebel	NLP + Anti-Gaming-Layer	E3
F10	Direkt-Antwortbarkeit (FAQPage, Tabellen)	Direkter Indikator	HTML-Parser + Schema-Probe	E1
F11	Off-Page-Authority (steuerbar)	Direkter Indikator	Link-Profil + Domain-Whitelists	E2
F12	Aktualitäts-Signale (datePublished, dateModified)	Direkter Indikator	JSON-LD-Parser	E1
Sekundäre Signale (max +10 Bonus)				
S1	Wikipedia-Eintrag mit Domain-Verweis (max +3)	Authority-Marker	Wikipedia-API	E1
S2	Branchenverbands-Mitgliedschaft (max +2)	Authority-Marker	Manuelle Whitelist	E2
S3	News-Coverage 12 Monate (max +2)	Authority-Marker	News-Aggregator-Probe	E2
S4	llms.txt vorhanden + valide (max +2)	Direkter Indikator	HTTP-Probe + Validator	E1
S5	Person-Schemas E-E-A-T-Tiefe (max +1)	Direkter Indikator	JSON-LD-Parser	E1

EVIDENZGRADE: DEFINITIONEN

E1 (Harte Fakten): direkt messbar, reproduzierbar, technisch eindeutig.
E2 (Fundierte Vermutungen): Korrelation in dokumentierten Cases plausibel, aber nicht kausal-deterministisch.
E3 (Explorativ): Proxy-Indikator, Konstrukt-Validität noch nicht abschließend geklärt; wird in der Wirkungs-Bibliothek weiter validiert.

ANHANG B

Glossar

BVR: Brand-Visibility-Rate
Sub-Metrik PS1. Anteil der Brand-Prompts, in denen die Marke vom LLM erwähnt wird.

CVR: Category-Visibility-Rate
Sub-Metrik PS2. Anteil der Category-/Core-Prompts (ohne Markenbezug), in denen die Marke spontan vom LLM erwähnt wird.

DiD (Difference-in-Differences)
Counterfactual-Methode aus der Ökonometrie. Trennt den kausalen Effekt einer Maßnahme vom allgemeinen Zeit-Trend, indem die Veränderung beim Treatment-Subjekt mit der Veränderung bei einer Kontrollgruppe (hier: Branchen-Referenzdomains) verglichen wird.

E-E-A-T
Experience, Expertise, Authoritativeness, Trustworthiness. Ursprünglich Google Quality Rater Guidelines, hier als Schema-Tiefe-Indikator für Autoren-Profile relevant.

F1–F12: Faktoren
Die zwölf gewichteten Mess-Faktoren des RS-Audits, gegliedert in drei Gruppen mit den Gewichten 25 %, 35 % und 40 %.

G1–G4: Gates
Die vier binären K.O.-Kriterien des RS-Audits. Wenn ein Gate gerissen ist, ist die KI-Verarbeitbarkeit der Website strukturell blockiert.

GEO: Generative Engine Optimization
Optimierung von Inhalten und Strukturen einer Website mit dem Ziel, in Antworten generativer KI-Systeme zitiert oder als Marke genannt zu werden.

MLC: Multi-LLM-Coverage
Sub-Metrik PS3. Anzahl der LLM-Systeme (0-4), in denen die Marke erscheint.

M1: RS-Audit
Mess-Modell für die strukturelle Empfangsbereitschaft der Website.

M2: PS-Tracking
Mess-Modell für das beobachtete Citation-Verhalten in vier LLM-Systemen.

M3: Wirkungsmessung
Mess-Modell für den kausalen Effekt einzelner Maßnahmen via DiD.

M4: Kontext-Map
Strategischer Rahmen mit fünf Dimensionen, der die Mess-Werte der anderen Modelle interpretiert.

PS: Performance-Profile
Beobachtetes Citation-Verhalten der Marke in KI-Antwort-Systemen, gemessen über fünf Sub-Metriken (BVR, CVR, MLC, CPQ, ASC).

RS: Readiness-Score
Strukturelle Empfangsbereitschaft einer Website für KI-Verarbeitung. Skala 0-100, bestehend aus 4 Gates, 12 Faktoren und 5 Signalen.

Wirkungs-Bibliothek
Sammlung dokumentierter Pre/Post-Cases mit Hypothesen, Auswertungen und Konfidenz- Stufen. Empirische Datenbasis für die laufende Validierung des Frameworks.

ANHANG C

Definitionsmatrix

MODELL	KONSTRUKT	MESS-METHODE	OUTPUT	ROLLE	EVIDENZ
M1	Empfangsbereitschaft	4 Gates + 12 Faktoren + 5 Signale	RS-Score 0–100	Diagnose-Modell	E1/E2 (E3 für F9)
M2	Beobachtetes Citation-Verhalten	CSV-Analyse aus LLM-Monitoring-Tool, 4 LLM-Systeme	5 Sub-Metriken (kein Composite)	Beobachtungs-Modell	E1 (mit Stochastik-Vorbehalt)
M3	Kausaler Effekt einer Maßnahme	Pre/Post + DiD vs Branchen-Referenzen	Effekt-Größe + Konfidenz	Diagnose-Modell	E2/E3
M4	Strategischer Kontext	5 Dimensionen aus Onboarding + Recherche	Lesart (2-4 Sätze) + Klassifikationen	Interpretations-Rahmen	E2

QUELLEN

Referenzen und weiterführende Literatur

Methodische Grundlagen

- **Card, D. & Krueger, A. B. (1994).** „Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." American Economic Review, 84(4), 772–793. Klassische Einführung der Difference-in-Differences-Methodik.
- **Angrist, J. D. & Pischke, J.-S. (2009).** Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. Kapitel 5 zu DiD und Counterfactual-Logik.
- **Manning, C. D., Raghavan, P. & Schütze, H. (2008).** Introduction to Information Retrieval. Cambridge University Press. Grundlagen der Indexierung und Retrieval-Bewertung.
- **Pearl, J. (2009).** Causality: Models, Reasoning, and Inference. 2. Auflage, Cambridge University Press. Kausalitäts-Theorie als Basis für M3.

Standards und Spezifikationen

- **Schema.org (2024).** Schema.org Vocabulary, Full Hierarchy. Online unter schema.org. Basis für F4-Faktor und JSON-LD-Implementierung.
- **Google (2024).** Core Web Vitals: Web Performance Metrics. Online unter web.dev/vitals. Basis für F1-Faktor.
- **llmstxt.org (2024).** llms.txt: Standard for LLM-Aware Robots Files. Basis für Signal S4.

Inter-Rater-Reliability und Reproduzierbarkeit

- **Cohen, J. (1960).** „A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement, 20(1), 37–46. Cohen's Kappa als Standard-Maß für Inter-Rater-Reliability.
- **Krippendorff, K. (2018).** Content Analysis: An Introduction to Its Methodology. 4. Auflage, SAGE Publications. Methodische Grundlagen für reproduzierbare Mess-Verfahren.

NLP und Sprach-Modellierung

- **spaCy (2024).** spaCy Models, de_core_news_lg (aktuelle Version der 3.x-Reihe). Online unter spacy.io/models/de. Basis für Anti-Gaming-NLP-Layer.
- **Bender, E. M. & Koller, A. (2020).** „Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." Proceedings of ACL 2020, 5185–5198. Konzeptuelle Grenzen von LLM-Output-Bewertung.

GEO/AI-Search-spezifische Diskussion

- **Otterly (2024–2026).** AI Search Visibility Tracking, Methoden-Dokumentation. Online unter otterly.ai. Aktuell verwendete Referenz-Implementation für M2 PS-Tracking — siehe Anhang E zu Mindest-Anforderungen und kompatiblen Alternativen.
- **SearchScore (2026).** „How to Measure and Track Your GEO Performance." Vergleichendes GEO-Scoring-Framework.

- **GenOptima (2026).** „Top 10 Generative AI Search Engine Optimization Agency Criteria for AEOaaS Readiness Score in 2026." Markt-Vergleich GEO-Agentur-Bewertung.

ANHANG E

Referenz-Implementation der LLM-Monitoring-Schicht

Das Framework ist **tool-agnostisch** konzipiert. Die methodische Beschreibung in §1–§13 und in den Anhängen A–C nennt bewusst keine konkreten Monitoring-Tools, sondern spricht generisch von „LLM-Monitoring-Tool", „Citation-Daten" oder „Monitoring-CSV". Damit lassen sich die fünf PS-Sub-Metriken (BVR, CVR, MLC, CPQ, ASC), der §9.5 Wettbewerber-Pool und das DiD-Counterfactual mit jedem Tool umsetzen, das die Mindest-Daten-Anforderungen erfüllt.

E.1 Mindest-Daten-Anforderungen an ein kompatibles Monitoring-Tool

Ein Tool ist mit dem Framework v3.3 kompatibel, wenn es die folgenden Daten als CSV-Export oder über eine API liefert:

ANFORDERUNG	BEGRÜNDUNG
Multi-LLM-Coverage über mindestens drei der vier im Framework getesteten Systeme (ChatGPT, Microsoft Copilot, Google AI Overview, Perplexity)	MLC-Sub-Metrik braucht mindestens drei Systeme für aussagekräftige Streuung
Tagesgenaue Citation-Häufigkeit pro Domain pro Prompt	DiD-Pre-Trend-Test (§5.4) und Pool-Stabilitäts-Filter benötigen Tages-Granularität
Konfigurierbare Wettbewerber-Liste (mindestens 5 Domains) für vergleichende Sichtbarkeitsmessung	§9.5 kunden-individueller Pool und Branchen-Authority-Whitelist
Konfigurierbare Prompt-Liste (mindestens 15 Prompts) mit Klassifizierung BRAND / CORE / ROTATION	Operative Promptliste-Erstellung ist auf 15 active Prompts kalibriert
Monatlicher CSV-Export mit Spalten: Datum, LLM-System, Prompt-ID, zitierte URL, Position in der Antwort	C3-Monatsreport-Pipeline ist auf dieses Schema kalibriert

E.2 Aktuell verwendete Referenz-Implementation

Otterly.ai ist die zum Stand v3.3 von Johannes Bopp GmbH operativ verwendete Referenz-Implementation. Otterly erfüllt alle Mindest-Anforderungen aus E.1 und ist seit Q1/2025 Datenquelle für M2-PS-Tracking sowie M3-Wirkungsmessung in allen aktiven Kundenprojekten. Die in §4.4 dokumentierten Mess-Metadaten-Blöcke nennen Otterly explizit als Tool-Bezeichnung, was die Replizierbarkeit der dokumentierten Cases sicherstellt.

E.3 Bekannte kompatible Alternativen

Stand 11. Mai 2026 sind die folgenden weiteren Tools bekannt, die die Mindest-Daten-Anforderungen erfüllen:

- **Peec.ai** — vergleichbarer Funktionsumfang, zusätzlich Citation-Gap-Analysis (zeigt Domains, in denen Wettbewerber zitiert sind aber das eigene Unternehmen nicht). Looker-Studio-Connector verfügbar. Pricing zum Stand Mai 2026 etwas höher als Otterly, mit Add-on-Logik für zusätzliche LLM-Systeme.
- **Profound** — US-fokussiertes Tool mit ähnlicher Datenstruktur, in DACH bisher weniger verbreitet.

Bei einem Wechsel oder zusätzlichen Tool-Einsatz ist eine Adapter-Schicht in der Pipeline notwendig (CSV-Format-Mapping pro Tool, geschätzter Aufwand 4–6 h Senior-Entwicklung). Die Methodik bleibt unverändert.

E.4 Begründung für die Tool-Agnostik

Die methodische Authority des Frameworks soll nicht von der Verfügbarkeit oder Pricing-Politik eines einzelnen kommerziellen Tools abhängen. Eine tool-agnostische Beschreibung (a) erleichtert externe Replikation und Peer-Review (Voraussetzung für die langfristig angestrebte Standardisierungs-Reife — siehe §1 Selbstpositionierung), (b) reduziert das Lock-in-Risiko für Anwender und (c) erlaubt es Beratungs-Empfehlungen, das je nach Kunden-Kontext passendste Tool zu nennen, ohne in einen Verkaufs-Konflikt zu geraten.

ANHANG F

PS-Tagesvariations-Studie (NEU v3.3.3)

Das Framework charakterisiert PS in §4 als semi-stochastischen Indikator mit Tagesvariation, die methodisch transparent gemacht werden muss. Diese Studie quantifiziert die Tagesvariation der fünf PS-Sub-Metriken empirisch über die aktuelle Wirkungs-Bibliothek- Stichprobe (Stand 12. Mai 2026).

F.1 Studien-Design

- **Datenquelle:** CSV-Export des LLM-Monitoring-Tools (siehe Anhang E zur aktuellen Referenz-Implementation), 30-Tage-Rolling-Daten der Pre- Maßnahme-Periode
- **Beobachtungs-Fenster:** 29.04.2026 bis 11.05.2026 (13 Tage)
- **Stichprobe:** n = 6 etablierte Kunden mit ≥ 12 Tagen Monitoring-Daten (Cases A bis F nach §9-Anonymisierung)
- **Ausgeschlossen:** 2 Kunden mit Tool-Setup < 4 Wochen alt (Case G mit nur 3 Tagen, plus ein Setup-Kunde mit 8 Tagen) — methodisch zu dünn für aussagekräftige Variations-Berechnung
- **Statistik:** Coefficient of Variation (CV%) pro Kunde pro Sub-Metrik, aggregiert als Mean / Median / Range über die Stichprobe

F.2 Methodik

Pro Kunde wurden täglich die fünf PS-Sub-Metriken aus den Detail-Daten des Monitoring-Tools berechnet (BVR aus BRAND-Prompts, CVR aus CORE-Prompts, MLC aus distinkten LLM-Systemen mit Brand- Citations, CPQ aus mittlerer Position eigener Brand-Citations, ASC aus distinkten Authority-Domains in den Domain-Categories Government/NGO, Education, News/Media und Encyclopedia).

Aus den Tageswerten wurde der Coefficient of Variation berechnet (CV% = Standardabweichung / Mittelwert × 100). Sub-Metriken mit konstantem Wert 0 wurden ausgeschlossen (kein definierter CV).

Die Studie ist vollständig reproduzierbar. Das Auswertungs-Skript (`ps_variation_study.py`), die anonymisierten CSV-Snapshots (Pilot-Case-Labels A–H) und das aggregierte JSON-Ergebnis sind als Bundle `PS-Variations-Studie_v3.3.5.zip` Bestandteil dieser Veröffentlichung (Zenodo-Datensatz, DOI siehe Zitierfeld).

F.3 Ergebnisse

SUB-METRIK	N KUNDEN	MEAN CV%	MEDIAN CV%	RANGE CV%	INTERPRETATION
BVR Brand-Visibility-Rate	5	23,5 %	11,2 %	4,9 – 69,1 %	Moderate Variation; Ausreißer bei Kunden mit niedriger Citation-Frequenz
CVR Category-Visibility-Rate	4	30,4 %	38,2 %	0,0 – 45,1 %	Höchste Variation; CORE-Prompts mit wenigen Brand-Citations sind volatil
MLC Multi-LLM-Coverage	6	13,5 %	3,5 %	0,0 – 57,0 %	Niedrige Median-Variation; bei niedriger Citation-Basis hohe Volatilität
CPQ Citation-Position-Quality	6	14,6 %	12,1 %	5,7 – 25,1 %	Stabilste Sub-Metrik; mittlere Position variiert wenig wenn Citations vorhanden
ASC Authority-Source-Coverage	6	15,3 %	14,3 %	8,7 – 26,0 %	Moderate Variation; Authority-Domain-Set ist über Tage vergleichsweise stabil

F.4 Konsequenzen für die PS-Interpretation

Die empirisch belegte Tagesvariation bestätigt die methodische Charakterisierung von PS als semi-stochastischen Indikator. Konkrete operative Konsequenzen:

1. **Einzeltageswerte sind nicht aussagekräftig.** Mean CV von 14–30 % über die meisten Sub-Metriken zeigt, dass tagesbasierte Aussagen („PS hat sich gestern verbessert") methodisch nicht belastbar sind.
2. **Wochenmittel oder Monatsmittel als Standard.** Für strategische Aussagen werden Sub-Metriken über mindestens 7 Tage gemittelt (idealerweise 30-Tage-Monatsdurchschnitt aus dem CSV-Rolling des Monitoring-Tools).
3. **Effekt-Größen-Schwellen mit Sicherheits-Abstand.** Eine M3-Auswertung setzt Mindest-Effekt-Größen, die über der gemessenen Tagesvariation liegen müssen (z. B. ΔCVR ≥ 10 % bei einer median CVR-Variation von 38 %, also ein Effekt nahe an der Median-Variation — die DiD-Analyse braucht hier zusätzliche Pre-Trend-Validierung in §5.4).
4. **Sub-Metrik-spezifische Konfidenz.** CPQ ist mit Median 12 % CV die stabilste Metrik und eignet sich für quartalsweise Trendaussagen; CVR mit Median 38 % CV eignet sich nur für Monats- oder Quartals-Mittel, nicht für Wochenwerte.

F.5 Limitationen

- **Stichprobengröße:** n = 6 Kunden ist klein. Cross-Branchen- Generalisierbarkeit ist eingeschränkt; die Studie wird mit jedem weiteren Pre-Snapshot aktualisiert.
- **Zeitfenster:** 13 Tage ist die untere Grenze für CV-Berechnung. Eine saubere ICC-Studie über 30+ Tage ist in der v4.0-Roadmap vorgesehen, geplant Q4/2026 nach n ≥ 15 longitudinalen Cases.
- **Sub-Metriken mit konstantem Nullwert** (z. B. wenn ein Kunde in keinem der Tage in CORE-Prompts erscheint) sind aus der CV-Berechnung ausgeschlossen, weil CV bei Mittelwert 0 nicht definiert ist. Diese Kunden zeigen aber strukturell hohe Konstanz auf niedrigem Niveau.
- **Bias-Möglichkeit:** Die Stichprobe enthält Kunden in unterschiedlichen Reife-Stufen (RS-Pre 33 bis 88). Eine Stratifizierung nach RS-Klasse ist erst mit n ≥ 10 pro Klasse statistisch sinnvoll.

Diese Studie adressiert Major Revision MR3 aus dem externen Methoden-Review (Perplexity, 11. Mai 2026) als erste empirische Etappe. Eine vollständige ICC-Studie (Inter-Rater-Reliabilität mit längerem Zeitfenster und stratifizierter Stichprobe) ist Hauptziel der Wirkungs-Bibliothek-Aufbauphase für v4.0.

KOLOPHON

Über dieses Dokument

Titel	GEO-Score Framework v3.3.5 · Whitepaper
Version	3.3.5
Stand	11. Mai 2026
Herausgeber	Johannes Bopp GmbH, Braunschweig
Marke	kmugeo, die GEO-Agentur für den DACH-Industriemittelstand
Autor	Tobias Ackermann (Head of GEO Operations & AI Visibility)
Lizenz	Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
Standardzitation	Ackermann, T. (2026). GEO-Score Framework v3.3.5. Johannes Bopp GmbH. Verfügbar unter kmugeo.de/geo-score-framework

Zitierfeld (NEU v3.3.1)

APA 7:

> Ackermann, T. (2026). *GEO-Score Framework* (Version 3.3.5) [Methodisches Whitepaper]. Johannes Bopp GmbH (kmugeo). Zenodo. https://doi.org/10.5281/zenodo.20137223

BibTeX:

```
@techreport{ackermann2026geo,
  author      = {Ackermann, Tobias},
  title       = {{GEO-Score Framework v3.3.5: Methodisches Whitepaper}},
  year        = {2026},
  month       = {May},
  institution = {Johannes Bopp GmbH},
  url         = {https://kmugeo.de/geo-score-framework},
  publisher   = {Zenodo},
  doi         = {10.5281/zenodo.20137223},
  url         = {https://doi.org/10.5281/zenodo.20137223},
  note        = {Creative Commons Attribution-ShareAlike 4.0 International}
}
```

Veröffentlicht via Zenodo am 12. Mai 2026 unter DOI `10.5281/zenodo.20137223`. Reproduzierbarkeits-Bundle (Skript + anonymisierte Daten) als ZIP im selben Datensatz.

Aktueller Status (v3.3.5) — publishing-ready

Final-Review-Response nach zweiter externer Methoden-Review (Perplexity, 12. Mai 2026, FINAL DECISION: ACCEPT). Vier punktuelle Verfeinerungen: (1) Inline-Verweis auf Anhang F im PS-Tracking-Kapitel mit Quantifizierung der Tagesvariation; (2) neue PS-Sub- Metrik-Stabilitäts-Matrix in Anhang A. 0.b mit CV-Werten aus Anhang F; (3) Quantifizierungs- Hinweis in Executive Summary Kernaussage 2; (4) Roadmap-Präzisierung ICC-Studie Q4/2026 nach n ≥ 15 longitudinalen Cases in Anhang F.5. Reviewer-Zitat: Mit Anhang F ist das Framework vollständig peer-review-tauglich für praxiswissenschaftliche Publikationen. Methodik unverändert.

Versionshistorie (kompakt)

VERSION	DATUM	ANLASS
v3.3.5	12.05.2026	Final-Review-Response (Perplexity ACCEPT): 4 Verfeinerungen + PS-Stabilitäts-Matrix
v3.3.4	12.05.2026	Konsistenz-Patch Tool-Agnostik in Anhang F + Pfad-Verweis
v3.3.3	12.05.2026	Empirie-Patch: Anhang F PS-Tagesvariations-Studie (n=6, Median CV 3,5–38,2 %)
v3.3.2	12.05.2026	Datenschutz-Patch (Anonymisierung DE) + Cases-Erweiterung 2→7 + Pre-Registration-Konvention
v3.3.1	11.05.2026	Quick-Wins nach Perplexity-Methoden-Review (5 Minor Revisions)
v3.3	11.05.2026	Sammel-Bump KW 19/20: Tool-Agnostik, B3.9-Pfad-Trennung, Pre-Trend-Test, Pool-Stabilitäts-Filter
v3.2.3	09.05.2026	Hybrid-Modell DiD-Datenquelle (kunden-individueller Pool §9.5 + Branchen-Whitelist)
v3.2.2	08.05.2026	Operative Mess-Mechanik (T+14/T+30 retrospektiv aus 30-Tage-Rolling)
v3.2.1	08.05.2026	Patch-Release nach drei Methoden-Review-Runden (Begriffs-Rollen, Sammelvorbehalts-Formel)

Vollständige Detail-Beschreibungen aller Bumps sowie die Pflege-Konvention für künftige Versionen in der mitveröffentlichten `CHANGELOG.md` dieses Datensatzes.

kmugeo · Eine Marke der Johannes Bopp GmbH