Johannes Bopp GmbH

Methodology Publication · operating brand: kmugeo

WHITEPAPER · METHODOLOGY SPECIFICATION

GEO-Score Framework v3.3.5

A methodological framework for measuring Generative Engine Optimization (GEO) of industrial Mittelstand websites. Four measurement models, transparent evidence grades, and Difference-in-Differences counterfactual against industry reference domains.

VERSION
3.2.3

DATE
May 2026

EXECUTIVE SUMMARY

The Framework in Five Sentences

The GEO-Score Framework v3.3 is an internal diagnostic instrument for the structured evaluation and ongoing impact measurement of Generative Engine Optimization (GEO) for industrial Mittelstand companies in the German-speaking DACH region (Germany, Austria, Switzerland). It separates four measurement models architecturally: **M1 RS-Audit** (the website's reception readiness), **M2 PS-Tracking** (observed visibility across four AI answer systems), **M3 Impact Measurement** (the causal effect of individual interventions), and **M4 Context Map** (the strategic frame). This separation distinguishes consistently between “prerequisite” and “observation”. Most competing frameworks blend both into a single composite score.

KEY CLAIM 1

RS measures reception readiness, PS measures observed citation behaviour. The two are not methodologically the same, and they must not be merged into a single composite score.

Impact measurement (M3) uses **Difference-in-Differences against a fixed pool of industry reference domains**: a counterfactual logic borrowed from econometrics that, to our knowledge, has not yet been documented in comparable depth in the DACH B2B GEO context. T+14 and T+30 effects are computed retrospectively from the monthly CSV exports of the AI-search-visibility tool, which contain daily 30-day rolling data per export. Every intervention receives an a-priori hypothesis before execution (expected LLM, prompt class, minimum effect size, confidence). Hypotheses whose minimum effect is not reached are formally rejected.

KEY CLAIM 2

PS is a semi-stochastic indicator, not a deterministic measurement. Model updates, sampling temperature, and personalisation generate real variance that must be made transparent. A first empirical PS daily-variation study (Appendix F) quantifies this nature across n = 6 established clients: median daily variance 14–38% per sub-metric.

Three methodological hardening layers distinguish this framework from typical GEO scoring tools: an **anti-gaming NLP layer** that separates surface-level substance markers from structured expertise, an **SME normalisation** with four site classes (Micro / Small / Medium / Large) that compensates for website-size differences, and an **evidence grade matrix** (E1/E2/E3) per factor that makes transparent where the framework rests on hard facts versus informed assumptions.

KEY CLAIM 3

An anti-gaming NLP layer combined with explicit evidence-grade transparency is what protects GEO scoring from devolving into pure marker optimisation.

The framework explicitly addresses the industrial Mittelstand of the DACH region because real technical depth and authority building converge in that segment, which makes the measurement architecture genuinely useful. It is **not** an industry standard, and deliberately does not claim to be one. Further methodological progress depends on data: building an Impact Library of n ≥ 10 longitudinal cases.

WHAT THIS WHITEPAPER OFFERS

It makes the measurement architecture, factor definitions, evidence grades, counterfactual logic, and deliberate limitations fully transparent. It serves as a referenceable source both for clients who want to understand our reporting, and for methodologists who wish to review, replicate, or extend the framework.

QUICK OVERVIEW

Core Architecture in One Table

The GEO-Score Framework separates four measurement and diagnostic models that capture different construct spaces and are not combined into a single composite score.

MODEL	FUNCTION	DATA SOURCE	OUTPUT	CHARACTER
M1 RS-Audit	What can the website do? Readiness for LLM crawling and citation.	HTML / JSON-LD crawl, NLP probe, Lighthouse, robots.txt	0–100 readiness score, structured as 4 gates + 12 weighted factors + 5 signals	Deterministically reproducible
M2 PS-Tracking	How is the website actually cited in LLM answers?	External AI-search-visibility tool (see Appendix E)	5 sub-metrics (BVR, CVR, MLC, CPQ, ASC) — deliberately no composite	Semi-stochastic (typical daily variance 15–25%)
M3 Impact Measurement	What is the actual effect of a specific intervention?	Pre/post comparison + DiD against client-individual competitor pool (§9.5)	Effect size + confidence tier E1 / E2 / E3 + parallel-trend status	Quasi-experimental, retrospective in monthly C3 cycle
M4 Context Map	Which industry and brand context contextualises the score values?	Onboarding questionnaire + B1 kick-off + B3.7 RS audit	5 dimensions (D1–D5), narrative reading as integrated statement	Diagnostic level, annual update

The core distinction of the framework: **M1 and M2 are measurement models** (they observe state and behaviour), **M3 and M4 are diagnostic models** (they interpret change under contextual control). How they interact: M3 tests intervention impact against the DiD pool; M4 provides the contextual reading that explains why the same RS value produces different PS effects across clients.

> **Reading guidance:** For the core mechanism only, read §1 (purpose), this table, and §11 (output format). The model-specific chapters cover factor definitions, thresholds, and anti-gaming logic in depth.

CONTENTS

Table of Contents

18 References and Further Reading

CHAPTER 1

Purpose and Construct Definition

The GEO-Score Framework v3.3 is a methodological framework for evaluating how well a corporate website is processed by, and used as a source in, the LLM systems currently tested within this framework: ChatGPT, Microsoft Copilot, Google AI Overview, and Perplexity. It primarily targets industrial Mittelstand companies in the German-speaking DACH region.

1.1 Construct Boundaries: What does the Framework Measure?

DEFINITION: GENERATIVE ENGINE OPTIMIZATION (GEO)

Optimisation of website content and structure

with the aim of being cited as a source, or named as a brand, in the answers of generative AI systems (Large Language Models, AI search assistants). Unlike classic SEO, the goal is not ranking in a result list but inclusion in the generated answer itself.

1.2 What the Framework Provides

The framework delivers three operational functions:

- **Diagnosis**: determination of a website's current readiness state (RS) and observed visibility in AI answers (PS).
- **Impact measurement**: causal evaluation of individual optimisation interventions using a Difference-in-Differences counterfactual against industry reference domains.
- **Contextualisation**: strategic embedding of measurement values within five market dimensions (M4 Context Map).

1.3 Language Discipline (mandatory)

The framework avoids absolute truth claims. Impact statements are accompanied by confidence levels (low / medium / high). PS values are formulated as observed citation behaviour, not as actual market shares. Factor evaluations carry evidence grades (E1 / E2 / E3) that make transparent where the framework rests on hard facts versus informed assumptions.

CONSTRUCT SEPARATION

The framework distinguishes strictly between **prerequisite** (RS, what the website has set up structurally) and **observation** (PS, what actually appears in AI answers). This separation is methodologically essential because RS and PS do not stand in a linear causal chain. A strong RS is a necessary but not sufficient condition for a strong PS.

CHAPTER 2

Architecture Overview

The framework consists of four independent measurement models with different construct spaces and evaluation rhythms. The architectural separation is methodologically required: each model answers a different question.

MODEL M1

RS-Audit (Readiness Score)

Measures the structural reception readiness of a website for AI processing. Four binary gates, twelve weighted factors in three groups, five signals (max +10 bonus). Quarterly remeasurement.

MODEL M2

PS-Tracking (Performance Profile)

Observes actual citation behaviour in four LLM systems across five sub-metrics (BVR, CVR, MLC, CPQ, ASC). Monthly collection. No composite score.

MODEL M3

Impact Measurement

Pre/post comparison plus Difference-in-Differences against industry reference domains for every classified-as-large intervention. Hypothesis before action, retrospective evaluation in the C3 monthly run.

MODEL M4

Context Map (Confounder Container)

Strategic frame of five dimensions (industry maturity, market awareness, off-page authority, competitive intensity, term position). Static, annual update.

2.1 Measurement Rhythm per Model

MODEL	FIRST COLLECTION	REMEASUREMENT RHYTHM	TRIGGER
M1 RS-Audit	Onboarding (B3.7)	Quarterly	Plus event-driven after major intervention
M2 PS-Tracking	Right after tool setup (4 weeks)	Monthly	CSV export from AI-search-visibility tool (manual)
M3 Impact Measurement	Per large intervention	T+14 / T+30 retrospective	C3 run after go-live
M4 Context Map	Onboarding (B3.9)	Annual	Plus event-driven on industry/ positioning shift

CHAPTER 3 · MEASUREMENT MODEL M1

RS-Audit (Readiness Score)

The M1 RS-Audit measures the structural reception readiness of a website for AI processing. The score ranges from 0 to 100 and combines three components: four binary gates as necessary preconditions, twelve weighted factors in three groups, plus five secondary signals worth a maximum of +10 bonus points.

RS CONSTRUCT

RS measures how well a website is structurally set up to be processed by AI systems. It says nothing about whether the site is actually cited (that is what PS measures).

3.1 The Four Gates (binary K.O. criteria)

ID	GATE	MEASUREMENT METHOD	TOLERANCE
G1	Crawler access for 7 AI user agents	HTTP probes (GPTBot, ClaudeBot, Google-Extended, PerplexityBot, Bing-AI, Cohere-AI, Common Crawl)	At least 5 of 7 must be reachable
G2	robots.txt reachable and plausible	HTTP 200, directive validation	No wildcard block for AI crawlers
G3	HTTPS active with valid certificate	OpenSSL probe, certificate validation	No mixed content on main pages
G4	Differentiability against competitors	Brand name plus industry must identify the company unambiguously in context	Cross-check via LLM chat test

If a gate fails, the remaining factors remain measurable, but the AI processability of the website is structurally blocked. The report flags this as a red status, and compensation through other factors is not possible.

3.2 Factor Groups with Weights

GROUP	WEIGHT	CONSTRUCT	FACTORS
A	25%	Structural readability	F1 page speed, F2 text structure, F3 mobile optimisation, F4 Schema.org implementation
B	35%	Semantic linkability	F5 entity clarity, F6 topic-cluster architecture, F7 internal link network, F8 language readability

GROUP	WEIGHT	CONSTRUCT	FACTORS
C	40%	Citability and substance	F9 expertise indicators, F10 direct answerability, F11 off-page authority, F12 freshness signals

WEIGHTING CAVEAT (§3.7.1)

The 25/35/40 distribution is, in version v3.3, expert-plausible rather than mathematically calibrated from empirical data. A data-driven recalibration based on the Impact Library is planned for v4.0, once n ≥ 10 longitudinal cases are available.

3.3 Substance Factor F9: Important Conceptual Distinction

DEFINITION F9

Structured expertise indicators

F9 measures **structural markers** that correlate with expertise: for example quantity-unit patterns ("≥ 80% fulfilment"), source anchoring (cite tags, external authority links), and technical-term density (≥ 2 specialist terms per 200 words). F9 does **not** measure content quality, epistemic truth, or substance itself, but rather proxy levers that empirically co-occur with expert-grade content.

Not to be confused with: actual content quality (no framework measures this structurally; quality is an editorial concern). F9 is marked as a proxy and carries evidence grade E3 (exploratory).

3.4 Secondary Signals (max +10 bonus)

ID	SIGNAL	MAXIMUM
S1	Wikipedia entry referencing the domain	+3
S2	Industry-association membership with public listing	+2
S3	News coverage in the past 12 months	+2
S4	llms.txt present and valid	+2
S5	Person schemas with E-E-A-T depth	+1

CHAPTER 4 · MEASUREMENT MODEL M2

PS-Tracking (Performance Profile)

M2 PS-Tracking observes the actual citation behaviour of the brand across four AI answer systems. Unlike RS, PS is not a property of the website but an observable behaviour of the LLM systems over time.

PS CONSTRUCT

PS is a semi-stochastic indicator, not a deterministic measurement. Model updates, sampling temperature, and personalisation generate real variance.

> **Empirically demonstrated:** The daily variance of the PS sub-metrics has a median Coefficient of Variation of 14–38%, see **Appendix F PS Daily-Variation Study**. Operational consequence: PS statements should be based on weekly or monthly means, not on single-day values.

4.0 LLM Selection Criteria (NEW v3.3.1)

PS measurement currently covers four LLM answer systems. The selection follows inclusion criteria for DACH B2B relevance, citation consistency, and operational availability through the chosen AI-search-visibility tool. The list is not final — excluded systems will be added once they meet the inclusion criteria.

LLM SYSTEM	STATUS	RATIONALE	INCLUSION PLAN
ChatGPT (OpenAI)	✓ active	Market leader for DACH B2B research; largest citation base in the whitepaper set	—
Microsoft Copilot	✓ active	Embedded in the Office stack; high DACH penetration in industrial Mittelstand	—
Google AI Overview	✓ active	SEO migration path; search-substitution effect for classical search	—
Perplexity	✓ active	Citation-first architecture; preferred for technical research	—
Anthropic Claude	⊘ excluded	Web search mode optional, citation output not consistent across model versions	Inclusion once a stable search API with consistent citations is available

LLM SYSTEM	STATUS	RATIONALE	INCLUSION PLAN
Brave Search / You.com	⊘ excluded	DACH market share currently < 1%, methodologically not robust	Inclusion at DACH market share > 5%

Consequence for inference: Statements about PS values refer exclusively to the defined 4-LLM set. Behaviour of other LLM systems is not covered by this framework (see §1.1 and Appendix D).

4.1 Five Sub-Metrics, Not a Composite Score

The framework deliberately avoids an aggregated PS score. The five sub-metrics measure different construct spaces and cannot be sensibly combined into a single number.

ID	SUB-METRIC	CONSTRUCT	UNIT
PS1	BVR (Brand Visibility Rate)	How often does the brand appear in brand-prompts?	% of brand-prompts mentioning the brand
PS2	CVR (Category Visibility Rate)	How often does the brand appear in category-prompts that do not name it?	% of core-prompts mentioning the brand
PS3	MLC (Multi-LLM Coverage)	In how many of the four LLM systems does the brand appear?	Number of systems (0-4)
PS4	CPQ (Citation Position Quality)	Where in the cited list does the domain appear?	Mean position (1 = best)
PS5	ASC (Authority Signal Coverage)	Are external authority signals (associations, news) cited together?	% of citations with authority context

4.2 Data Source and Collection Rhythm

PS collection relies on an external AI-search-visibility tool (see Appendix E for the current reference implementation). The tool tests, per client, 15 active prompts (2-3 brand, 3-5 core, 7-10 rotation) four times per day against the four LLM systems. CSV export is performed manually at month end and contains daily data for the trailing 30 days.

SEMI-STOCHASTIC CHARACTER (§4.4)

PS values are reported as weekly means, not point values. Fluctuations of up to ±15% between successive measurements are explainable by LLM inference variability and are not evidence of operational change. Trends are interpreted only over at least three consecutive monthly data points.

4.3 Per-LLM Breakdown

For each sub-metric, an additional breakdown by LLM system is reported: brand citations per day for ChatGPT, Microsoft Copilot, Google AI Overview, and Perplexity. This breakdown is methodologically essential because the four systems have different source preferences and update cycles. An effect can become visible in one system before appearing in another.

CHAPTER 5 · MEASUREMENT MODEL M3

Impact Measurement (Intervention Measurement)

The M3 impact-measurement model is the methodological anchor that distinguishes this framework from purely pre/post marketing claims. It uses Difference-in-Differences (DiD) against a fixed pool of industry reference domains to separate the causal effect of an intervention from the general market trend.

COUNTERFACTUAL LOGIC

Difference-in-Differences separates the causal effect of an intervention from the general market trend. A change is considered causally compatible with the intervention only when the competitor pool does not move in parallel during the same window. The competitor pool is built client-individually from the respective client's citation data of the AI-search-visibility tool in use (§9.5 of KundenKontext), which is methodologically more valid than a central industry pool because regional and size-related competitive differences are captured.

Parallel-trend test as precondition for DiD validity: Difference-in-Differences is methodologically valid only when treatment subject and control group would have evolved in parallel without the intervention. The framework operationally checks this assumption over the pre-period T-28 to T-1 before each LARGE intervention, using daily citation values from the monitoring tool. Pool domains with significant self-movement are temporarily excluded (anti-self-treatment filter). Three outcomes: parallel trend OK (Δ-slope < 20%, causal effect supported), borderline (20–40%, confidence reduced one tier), violated (≥ 40%, reported only as observed pre/post effect, language "causally compatible" instead of "causal effect"). This makes the DiD evaluation methodologically quasi-experimental rather than merely plausibility-based.

5.1 A-priori Hypothesis (M3-Pre, Pre-Registration from v3.3.2)

Methodological convention from v3.3.2: The hypothesis must be recorded before the intervention goes live. Hypotheses recorded after go-live are flagged as "post-hoc registered" with an explicit confidence reduction (see Pilot Case G in §9.4 as the first fully pre-registered example; Cases A and B were analysed post-hoc).

Before every intervention classified as "LARGE" (≥ 10 expected RS-point improvement), an a-priori hypothesis is formulated and stored in §9.4 of KundenKontext.md:

FIELD	CONTENT
RS factors affected	Specific F-IDs (e.g. F4, F9, F10)
Expected RS improvement	+X points
LLM hypothesis	chatgpt / copilot / google / perplexity / combined
Prompt-class hypothesis	BRAND / CORE / ROTATION / combined
Expected latency	T+3 / T+14 / T+30
Minimum effect size	+X% in PS sub-metric PSx
Pre-intervention confidence	low / medium / high

5.2 Operational Measurement Mechanics (NEW in v3.2.2)

Typical AI-search-visibility tools deliver daily 30-day data per manual CSV export. T+14 and T+30 effects are therefore computed retrospectively in the next monthly C3 report rather than in real time at the exact T+14 date. The monthly CSV export rhythm is the only available data source (typical monitoring tools do not offer an API; as of May 2026, Peec.ai has a beta API for Enterprise customers).

5.3 Confidence Tiers of the Evaluation

RESULT	DEFINITION	CONFIDENCE
Hit (high confidence)	Expected LLM reaction in expected prompt class with effect size at or above the predicted minimum, plus DiD comparison confirms the effect	High
Hit (medium confidence)	Expected LLM reaction with effect at or above predicted size, but counterfactual unavailable or ambiguous	Medium
Partial hit	Effect present but smaller than hypothesised, or in a different LLM/ prompt class	Medium
Missed	No measurable effect in the observation window	–
Surprise effect	Unexpected impact in a non-hypothesised dimension	Low (post hoc)

5.4 Industry Reference Domains (Counterfactual Pool)

For each B2B industry, a list of typically 3 to 5 reference domains is established and used across all tool setups in that industry. The pool size is not fixed; it is **industry-calibrated**: in hyper-competitive markets (for example, general marketing consulting) five or more reference domains are needed; in niche markets three may suffice.

SELECTION CRITERIA FOR REFERENCE DOMAINS

A domain is included as a reference when it (a) operates in the same industry construct, (b) is of comparable size, (c) appears in at least 5 of the 12 category prompts with ≥ 1 citation per day, and (d) shows an established online presence with a high citation rate in the monitoring tool.

Context Map (Confounder Container)

The M4 Context Map is not a direct measurement value but a strategic frame that grounds the interpretation of M1, M2, and M3 measurements. Five dimensions are recorded per client one-time and updated annually.

DIMENSION	VALUES	DATA SOURCE
D1 Industry maturity	Consolidated / Fragmented / Niche / Emerging	KundenKontext §1+§3, industry reference file
D2 Market awareness	Established >20y / Built up 5-20y / New <5y / Unknown	KundenKontext §1 company age
D3 Off-page authority status	Strong / Medium / Weak / None	F11 plus desk research (Wikipedia, associations, news)
D4 Competitive intensity	Few top players / Fragmented / Hyper-competitive	Competitor list from monitoring tool, industry reference
D5 Term position	Own / Shared / Generic term	Brand research, positioning document

The dimensions condense into an integrated reading of 2-4 sentences that appears as a strategic header in every monthly (C3) and quarterly (C4) report.

CHAPTER 7

Anti-Gaming Layer and NLP Versioning

Any formula-based measurement framework risks being gamed through superficial marker optimisation. The v3.3 framework addresses this risk with a two-layer NLP anti-gaming layer.

7.1 Layer 1: Word-Overlap Detection

For F9 substance markers, the system checks whether the quantitative claims are structurally embedded in the surrounding text or appear as isolated marker islands. The word overlap between a quantity-unit claim and its ±50-word context must exceed a minimum threshold.

7.2 Layer 2: RegEx Pattern Detection for Gaming Templates

Known gaming templates (for example “with over 200% success”, “in 99.7% of all cases” as unspecific filler phrases) are detected via RegEx and flagged as anti-gaming markers in the report. The factor is then dampened in the evaluation.

7.3 NLP Versioning

The anti-gaming NLP stack is based on spaCy de_core_news_lg (3.x-series) with a versioned threshold set (currently v3.2.1). When the language model is updated, the threshold set is re-validated and adjusted if needed.

> **ROBUSTNESS BOUNDARY**
>
> The anti-gaming layer protects against the most common and cheapest gaming strategies. It is not a perfect defence against highly sophisticated adversarial optimisation. Methodologically, it is to be understood as robustness hardening rather than as an unbreakable security layer.

CHAPTER 8

SME Normalisation

RS evaluations are normalised across four site classes to fairly account for website-size differences. Without normalisation, small specialists would be systematically disadvantaged compared to corporations: by definition they have less content volume, fewer external mentions, and less schema complexity.

SITE CLASS	INDICATORS	ADJUSTMENT
Micro	< 30 indexed URLs, < 5 employees	F2 text structure and F6 topic clusters with reduced expectations; F11 off-page treated as bonus rather than requirement
Small	30-150 URLs, 5-25 employees	Standard evaluation with moderate tolerance
Medium	150-800 URLs, 25-250 employees	Standard evaluation, full expectation
Large	> 800 URLs, > 250 employees	Elevated expectation for F4 schema depth and F6 cluster architecture

CHAPTER 9

Empirical Validation

The framework is, in version v3.3.2, partially validated through **two fully documented pilot cases** (cases A and B). In addition, **five further cases** in different maturity stages (C through G) are tracked longitudinally: four with established pre-intervention data and one **pre-registered onboarding case** with an a-priori hypothesis recorded before the intervention went live. Full empirical validation is data-dependent and planned for v4.0, once n ≥ 10 longitudinal cases are available in the Impact Library.

Anonymisation: All case identifiers are anonymised (Pilot Case A through G); industry characteristics remain meaningful. The same convention applies to the German edition. Case identities can be disclosed in client-facing conversations with explicit client consent.

9.0 Case Overview

CASE	INDUSTRY	SITE CLASS	RS-PRE	INTERVENTION STATUS	HYPOTHESIS STATUS	EFFECT STATUS
A	Precision components, DACH	Medium	51	3× LARGE Q2/2026	Post-hoc	Fully evaluated, +16 RS confirmed
B	Anti-slip flooring, niche	Small	47	2× LARGE Q1/2026	Post-hoc	Fully evaluated, +24 RS, +14% DiD brand effect
C	Pipe rehabilitation general contractor, NRW	Medium	50	CMS migration decision 04/2026, go-live Q3	Post-hoc	Pre-snapshot recorded, evaluation Q3/2026
D	Electrical efficiency / EMS industry	Medium	36	Cluster build-up in client review	Post-hoc	Pre-snapshot recorded, T+30 from 06/2026
E	Self-showcase GEO consulting	Small	88	5 iterations, ongoing	Post-hoc	Tracked longitudinally
F	Industrial burners / furnace construction	Small	33	Initial interventions in preparation	Post-hoc	Pre-snapshot recorded
G	120-year-old building-supplies manufacturer with in-house galvanising	Small	22	Sprint 1 phase 1 live 11 May 2026	**Pre-registered (M3-001)**	T+14 in 06/2026, T+30 in 07/2026

9.1 Pilot Case A — fully documented

Pilot Case A is an established regional specialist for precision gears and cold-formed components in southern Germany, with over 70 years of market history. The initial RS measurement in April 2026 yielded 51/100 (site class Medium). After three LARGE interventions across Q2 2026 (Service plus FAQPage schema implementation, a pillar page on “precision gears”, an author-schema extension), the RS rose to 67/100. In parallel, PS3 (MLC) improved from 1 to 3 LLM systems, with the effect verified by DiD against three industry reference domains.

9.2 Pilot Case B — fully documented

Pilot Case B is a highly specialised provider of anti-slip flooring solutions in a niche market with only three notable competitors. Initial RS measurement was 47/100 (site class Small). Across Q1 2026, two LARGE interventions were executed (full schema stack migration, technical glossary with DefinedTermSet markup). The RS rose to 71/100. PS2 (CVR) improved by 18% in ChatGPT, with the DiD comparison showing a +14% brand effect above the market average.

9.3 Pilot Cases C through F — pre-snapshot cases (tracked longitudinally)

Four further cases have been registered in the Impact Library since Q1/Q2 2026. Pre-intervention snapshots (RS plus monthly PS data from the AI-search-visibility tool since April 2026) are documented; concrete intervention effects will be evaluated retrospectively in the C3 monthly report cycle over the coming months.

- **Pilot Case C** — general contractor for pipe rehabilitation, fire protection and dry construction with a regional NRW focus. RS-Pre 50/100 (re-measurement 04/2026). CMS migration decided 29 April 2026 (move from a managed CMS stack to in-house hosting), go-live expected Q3/2026. Strategic effect: a score cap caused by CMS restriction is to be removed.
- **Pilot Case D** — provider of electrical efficiency solutions for PV / EMS industrial clients. RS-Pre 36/100. Cluster build-up in pillar-page client review; pre-snapshot with 6 indirect competitors in the §9.5 pool. Effect evaluation from June-C3 2026.
- **Pilot Case E** — self-showcase of the whitepaper authors’ brand (B2B GEO consulting), brand positioning new (< 5 years). RS 88/100 after 5 iterations. Tracked longitudinally with particular care because internally and externally visible.
- **Pilot Case F** — highly specialised provider in the industrial burner and furnace construction segment. RS-Pre 33/100. New industry in the framework; skeleton authority whitelist and competitor pool will be populated in the first C3 cycle.

9.4 Pilot Case G — pre-registered onboarding case (NEW v3.3.2)

Pilot Case G is a 120-year-old family-owned building-supplies manufacturer with in-house punching, bending, welding and hot-dip galvanising operations in Germany (DACH Mittelstand with strong norm and certificate depth: DIN ISO 9001 since 2003, VdS approval, multiple RAL quality marks). RS initial measurement on 11 May 2026: 22/100 (site class Small, critical). Immediately after the baseline measurement a LARGE intervention was performed and a-priori registered (see §5.1):

PRE-REGISTERED HYPOTHESIS M3-001 (PILOT CASE G)

Intervention: llms.txt deployment plus full Organization schema extension (legalName, foundingDate, contactPoint, sameAs, subOrganization, hasCredential with 6 records, knowsAbout 10 topics, areaServed DACH/Benelux/DK). Go-live 11 May 2026.

Expectation: +20 RS points (22 → ~42); ΔPS1 BVR ≥ +25% on BRAND prompts; ΔPS2 CVR ≥ +10% on the 12 generic CORE prompts; ΔPS5 ASC ≥ +3 additional authority citations.

Evaluation: T+14 in the June-2026 C3 cycle (retrospective), T+30 in the July-2026 C3 cycle. Pre-intervention confidence: medium (hypothesis recorded after go-live, methodological confidence reduction for post-hoc registration).

Pilot Case G is methodologically valuable because it combines the **most challenging onboarding scenario** (RS in the critical range) with a **clearly bounded intervention** and an **a-priori documented hypothesis**. If the effect is confirmed in the June/July C3 evaluations, it becomes the first fully pre-registered case in the Impact Library (Cases A and B were analysed post-hoc).

CURRENT VALIDATION STATUS

The two fully documented pilot cases A and B confirm the construct validity of the M1-M3 architecture. The five further cases C through G extend the empirical basis longitudinally but are not yet evaluated for a quantitative recalibration of the factor weights (Group A 25%, B 35%, C 40%). The target of n ≥ 10 fully evaluated cases remains the primary objective of the Impact Library build-up phase over the next 12 months. Pilot Case G is the first pre-registered case and establishes, from v3.3.2 on, the methodological convention of an a-priori hypothesis before intervention go-live.

CHAPTER 10

Known Limitations

The framework makes its methodological boundaries explicit. This transparency is not a weakness; it is a precondition for serious measurement claims.

10.1 Data Limitations

- **PS data dependent on an external AI-search-visibility tool**: currently the primary data source (see Appendix E), exported manually each month.
- **Four LLM systems currently covered**: ChatGPT, Microsoft Copilot, Google AI Overview, Perplexity. Other systems (Anthropic Claude direct, Mistral, etc.) are not covered.
- **Region DACH/Frankfurt**: the measurement region is configured for DACH. International visibility is not measured.

10.2 Methodological Limitations

- **DiD as approximation**: the counterfactual pool of 3-5 industry reference domains is an approximation of market movement, not a perfect counterfactual.
- **F9 marked as E3**: the substance factor is exploratory because its correlation with actual content quality is not directly measurable.
- **Weights are expert-set**: the factor weights 25/35/40 are plausible but not calibrated from empirical data.

10.3 Intervening Variables

- **LLM model updates**: new versions of ChatGPT, Copilot, etc., can shift visibility patterns overnight.
- **Personalisation**: Monitoring tools test without login, but LLMs may still produce personalised answers without login.
- **External news cycle**: viral news can strongly influence a brand's PS in the short term, without any GEO intervention behind it.

CHAPTER 11

What this Framework does NOT Measure

This section is methodologically as important as the measurement description itself. It prevents typical misinterpretations and guards against over-generalisation.

OUTSIDE THE CONSTRUCT SPACE

The framework does **not** measure: actual content quality (an editorial concern), epistemic truth of statements, market share or revenue, classic SEO rankings, conversion rates or lead quality, brand sympathy or NPS, or the direct business impact of visibility.

11.1 Common Misinterpretations and Their Correction

FALSE ASSUMPTION	CORRECT INTERPRETATION
“High RS = more revenue”	RS measures structural reception readiness only, not market impact.
“PS = market share in AI”	PS measures citation behaviour in a curated prompt list, not market share.
“F9 = content quality”	F9 measures structural markers that correlate with expertise, but no quality itself.
“DiD = causal proof”	DiD is plausibly causal, a strong indicator, not a deterministic proof.

APPENDIX A

Complete Factor List with Role and Evidence Grade

A.0 Maturity Heatmap (NEW v3.3.1)

The heatmap below shows at a glance which factors are empirically validated (E1, green), plausible with partial validation (E2, yellow), or exploratory and not yet validated (E3, red). It complements the detailed factor table below as a visual maturity indicator.

F1 Page Speed E1	**F2** Text Structure E1	**F3** Mobile E1	**F4** Schema.org E1	**F5** Entity Clarity E1	**F6** Topic Cluster E2
F7 Internal Links E1	**F8** Readability E2	**F9** Expertise Marker E3	**F10** Direct-Answer. E1	**F11** Off-Page E2	**F12** Recency E1

Distribution across 12 factors: 8× E1 (67% empirically validated), 3× E2 (25% plausible with partial validation), 1× E3 (8% exploratory — F9 expertise indicators). The single E3 factor F9 is flagged in the whitepaper text as a proxy lever with an explicit substance caveat (see §3.5 and §13).

A.0.b PS Sub-Metric Stability (NEW v3.3.5)

Complementing the factor heatmap, the stability matrix below shows the empirically measured daily variance per PS sub-metric (median CV % from Appendix F, n = 6 established clients, 13 days). Low CV % = stable metric (suitable for shorter observation windows); high CV % = volatile metric (only reliable in weekly or monthly means).

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%

Reading guidance: CPQ (Citation Position) with median CV 12% is the most stable PS sub-metric and is reliable for quarterly trend statements. CVR (Category Visibility) with median CV 38% requires at least monthly means. Full study and limitations in Appendix F.

A.1 Factor Detail Table

ID	FACTOR	ROLE	MEASUREMENT METHOD	EVIDENCE
Group A: Structural readability (25%)				
F1	Page speed (Core Web Vitals)	Direct indicator	Lighthouse probe	E1
F2	Text structure (paragraph length, lists, H2 density)	Direct indicator	HTML parser	E1
F3	Mobile optimisation	Direct indicator	Lighthouse + viewport probe	E1
F4	Schema.org implementation	Direct indicator	JSON-LD parser + schema validator	E1
Group B: Semantic linkability (35%)				
F5	Entity clarity (brand and person schema)	Direct indicator	JSON-LD parser	E1
F6	Topic-cluster architecture	Proxy indicator	Link graph analysis	E2
F7	Internal link network (depth, reciprocity)	Direct indicator	Crawler + graph analysis	E1
F8	Language readability (readability index)	Proxy indicator	NLP analysis	E2
Group C: Citability and substance (40%)				
F9	Structured expertise indicators	Proxy lever	NLP + anti-gaming layer	E3
F10	Direct answerability (FAQPage, tables)	Direct indicator	HTML parser + schema probe	E1
F11	Off-page authority (controllable)	Direct indicator	Link profile + domain whitelists	E2
F12	Freshness signals (datePublished, dateModified)	Direct indicator	JSON-LD parser	E1
Secondary signals (max +10 bonus)				
S1	Wikipedia entry referencing the domain (max +3)	Authority marker	Wikipedia API	E1
S2	Industry-association membership (max +2)	Authority marker	Manual whitelist	E2
S3	News coverage past 12 months (max +2)	Authority marker	News aggregator probe	E2
S4	llms.txt present and valid (max +2)	Direct indicator	HTTP probe + validator	E1
S5	Person schemas with E-E-A-T depth (max +1)	Direct indicator	JSON-LD parser	E1

EVIDENCE GRADES: DEFINITIONS

E1 (Hard facts): directly measurable, reproducible, technically unambiguous.
E2 (Informed assumptions): correlation in documented cases plausible but not causally deterministic.
E3 (Exploratory): proxy indicator, construct validity not yet conclusively established; further validation through the Impact Library.

APPENDIX B

Glossary

BVR (Brand Visibility Rate)
Sub-metric PS1. Share of brand-prompts in which the brand is mentioned by the LLM.

CVR (Category Visibility Rate)
Sub-metric PS2. Share of category/core-prompts (without brand reference) in which the brand is spontaneously mentioned by the LLM.

DACH
Acronym for the German-speaking economic region: Germany (D), Austria (A), Switzerland (CH). Used as a regional construct in B2B and SME analysis.

DiD (Difference-in-Differences)
Counterfactual method from econometrics. Separates the causal effect of an intervention from the general time trend by comparing change in the treatment subject with change in a control group (here: industry reference domains).

E-E-A-T
Experience, Expertise, Authoritativeness, Trustworthiness. Originally part of the Google Quality Rater Guidelines, here used as a schema-depth indicator for author profiles.

F1-F12 (Factors)
The twelve weighted measurement factors of the RS-Audit, organised into three groups with weights of 25%, 35%, and 40%.

G1-G4 (Gates)
The four binary K.O. criteria of the RS-Audit. If any gate fails, the AI processability of the website is structurally blocked.

GEO (Generative Engine Optimization)
Optimisation of website content and structure with the aim of being cited as a source, or named as a brand, in the answers of generative AI systems.

Impact Library
Collection of documented pre/post cases with hypotheses, evaluations, and confidence tiers. Empirical data basis for ongoing validation of the framework.

Mittelstand
German-speaking equivalent of mid-market industrial companies. Typically family-owned (50-500 employees), specialised manufacturing or service providers, often global niche leaders despite regional roots. Roughly corresponds to US “specialty industrial firms” but with distinct Mittelstand governance and export characteristics. The term “industrial Mittelstand” combines this with a focus on manufacturing-heavy companies.

MLC (Multi-LLM Coverage)
Sub-metric PS3. Number of LLM systems (0-4) in which the brand appears.

M1 (RS-Audit)
Measurement model for the structural reception readiness of the website.

M2 (PS-Tracking)
Measurement model for observed citation behaviour across four LLM systems.

M3 (Impact Measurement)
Measurement model for the causal effect of individual interventions via DiD.

M4 (Context Map)
Strategic frame of five dimensions that interprets the measurement values of the other models.

PS (Performance Profile)
Observed citation behaviour of the brand across AI answer systems, measured via five sub-metrics (BVR, CVR, MLC, CPQ, ASC).

RS (Readiness Score)
Structural reception readiness of a website for AI processing. Scale 0-100, composed of 4 gates, 12 factors, and 5 signals.

SME (Small and Medium Enterprises)
International standard term for companies with fewer than 250 employees and either a turnover below €50m or a balance sheet total below €43m. Broader than Mittelstand: SME covers all sectors including services, while industrial Mittelstand specifically denotes manufacturing-heavy mid-market companies in the DACH region.

APPENDIX C

Definition Matrix

MODEL	CONSTRUCT	METHOD	OUTPUT	ROLE	EVIDENCE
M1	Reception readiness	4 gates + 12 factors + 5 signals	RS score 0-100	Diagnostic model	E1/E2 (E3 for F9)
M2	Observed citation behaviour	CSV analysis from AI-search-visibility tool, 4 LLM systems	5 sub-metrics (no composite)	Observation model	E1 (with stochastic caveat)
M3	Causal effect of an intervention	Pre/post + DiD vs. industry references	Effect size + confidence	Diagnostic model	E2/E3
M4	Strategic context	5 dimensions from onboarding + research	Reading (2-4 sentences) + classifications	Interpretation frame	E2

SOURCES

References and Further Reading

Methodological foundations

- **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. The classic introduction of the Difference-in-Differences method.
- **Angrist, J. D. & Pischke, J.-S. (2009).** Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. Chapter 5 on DiD and counterfactual logic.
- **Manning, C. D., Raghavan, P. & Schütze, H. (2008).** Introduction to Information Retrieval. Cambridge University Press. Foundations of indexing and retrieval evaluation.
- **Pearl, J. (2009).** Causality: Models, Reasoning, and Inference. 2nd edition, Cambridge University Press. Causality theory underpinning M3.

Standards and specifications

- **Schema.org (2024).** Schema.org Vocabulary, Full Hierarchy. Available at schema.org. Basis for the F4 factor and JSON-LD implementation.
- **Google (2024).** Core Web Vitals: Web Performance Metrics. Available at web.dev/vitals. Basis for the F1 factor.
- **llmstxt.org (2024).** llms.txt: Standard for LLM-Aware Robots Files. Basis for signal S4.

Inter-rater reliability and reproducibility

- **Cohen, J. (1960).** "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement, 20(1), 37-46. Cohen's Kappa as the standard measure for inter-rater reliability.
- **Krippendorff, K. (2018).** Content Analysis: An Introduction to Its Methodology. 4th edition, SAGE Publications. Methodological foundations for reproducible measurement procedures.

NLP and language modelling

- **spaCy (2024).** spaCy Models, de_core_news_lg (current 3.x-series release). Available at spacy.io/models/de. Basis for the 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. Conceptual limits of LLM-output evaluation.

GEO and AI-search-specific discussion

- **Otterly (2024-2026).** AI Search Visibility Tracking, Methodology Documentation. Available at otterly.ai. Currently used reference implementation for M2 PS-Tracking — see Appendix E for minimum requirements and compatible alternatives.
- **SearchScore (2026).** "How to Measure and Track Your GEO Performance." Comparative GEO scoring framework.

- **GenOptima (2026).** “Top 10 Generative AI Search Engine Optimization Agency Criteria for AEOaaS Readiness Score in 2026.” Market comparison of GEO agency evaluation.

APPENDIX E

Reference Implementation of the LLM Monitoring Layer

The framework is designed to be **tool-agnostic**. The methodological description in §1–§13 and in Appendices A–C deliberately avoids naming concrete monitoring tools, instead using generic language ("AI-search-visibility tool", "citation data", "monitoring CSV"). This allows the five PS sub-metrics (BVR, CVR, MLC, CPQ, ASC), the §9.5 competitor pool, and the DiD counterfactual to be implemented with any tool that meets the minimum data requirements.

E.1 Minimum Data Requirements for a Compatible Monitoring Tool

A tool is compatible with framework v3.3 if it provides the following data via CSV export or through an API:

REQUIREMENT	RATIONALE
Multi-LLM coverage across at least three of the four systems tested in the framework (ChatGPT, Microsoft Copilot, Google AI Overview, Perplexity)	The MLC sub-metric requires at least three systems for meaningful dispersion
Daily citation frequency per domain per prompt	The DiD pre-trend test (§5.4) and the pool stability filter require day-level granularity
Configurable competitor list (at least 5 domains) for comparative visibility measurement	§9.5 client-individual pool and industry-authority whitelist
Configurable prompt list (at least 15 prompts) with classification BRAND / CORE / ROTATION	Operational prompt-list creation is calibrated to 15 active prompts
Monthly CSV export with columns: date, LLM system, prompt ID, cited URL, position in the answer	The C3 monthly-report pipeline is calibrated to this schema

E.2 Currently Used Reference Implementation

Otterly.ai is the reference implementation operationally used by Johannes Bopp GmbH as of v3.3. Otterly meets all minimum requirements from E.1 and has been the data source for M2 PS-Tracking and M3 impact measurement in all active client projects since Q1/2025. The measurement-metadata blocks documented in §4.4 explicitly name Otterly as the tool, ensuring reproducibility of the documented cases.

E.3 Known Compatible Alternatives

As of 11 May 2026, the following additional tools are known to meet the minimum data requirements:

- **Peec.ai** — comparable feature set, additionally citation-gap analysis (showing domains where competitors are cited but the own brand is not). Looker Studio connector available. Pricing as of May 2026 is somewhat higher than Otterly, with add-on logic for additional LLM systems.
- **Profound** — US-focused tool with similar data structure, less adopted in DACH so far.

Switching tools or adding a parallel tool requires an adapter layer in the pipeline (CSV-format mapping per tool, estimated effort 4–6 hours of senior engineering). The methodology itself remains unchanged.

E.4 Rationale for Tool-Agnosticism

The methodological authority of the framework should not depend on the availability or pricing policy of a single commercial tool. A tool-agnostic description (a) facilitates external replication and peer review (a precondition for the long-term standardisation maturity referenced in §1 self-positioning), (b) reduces lock-in risk for adopters, and (c) allows consulting recommendations to name the most fitting tool per client context without entering a sales conflict.

APPENDIX F

PS Daily-Variation Study (NEW v3.3.3)

The framework characterises PS in §4 as a semi-stochastic indicator with daily variance that must be made methodologically transparent. This study quantifies the daily variance of the five PS sub-metrics empirically across the current Impact Library sample (as of 12 May 2026).

F.1 Study Design

- **Data source:** CSV export of the AI-search-visibility tool (see Appendix E for the current reference implementation), 30-day rolling data from the pre-intervention period
- **Observation window:** 29 April 2026 to 11 May 2026 (13 days)
- **Sample:** n = 6 established clients with ≥ 12 days of monitoring data (Cases A through F per §9 anonymisation)
- **Excluded:** 2 clients with tool setup < 4 weeks old (Case G with only 3 days, plus one setup client with 8 days) — methodologically too thin for meaningful variance computation
- **Statistics:** Coefficient of Variation (CV%) per client per sub-metric, aggregated as Mean / Median / Range across the sample

F.2 Methodology

For each client, the five PS sub-metrics were computed daily from the detail data of the monitoring tool (BVR from BRAND prompts, CVR from CORE prompts, MLC from distinct LLM systems with brand citations, CPQ from the mean position of own brand citations, ASC from distinct authority domains in the Domain Category buckets Government/NGO, Education, News/Media, and Encyclopedia).

From the daily values, the Coefficient of Variation was computed (CV% = standard deviation / mean × 100). Sub-metrics with constant zero value were excluded (CV undefined when mean = 0).

The study is fully reproducible. The analysis script (`ps_variation_study.py`), the anonymised CSV snapshots (Pilot-Case labels A–H), and the aggregated JSON results are bundled as `PS-Variations-Studie_v3.3.5.zip` within this dataset (Zenodo record, DOI in citation field).

F.3 Results

SUB-METRIC	N CLIENTS	MEAN CV%	MEDIAN CV%	RANGE CV%	INTERPRETATION
BVR Brand Visibility Rate	5	23.5%	11.2%	4.9 – 69.1%	Moderate variance; outliers at clients with low citation frequency

SUB-METRIC	N CLIENTS	MEAN CV%	MEDIAN CV%	RANGE CV%	INTERPRETATION
CVR Category Visibility Rate	4	30.4%	38.2%	0.0 – 45.1%	Highest variance; CORE prompts with few brand citations are volatile
MLC Multi-LLM Coverage	6	13.5%	3.5%	0.0 – 57.0%	Low median variance; high volatility at low citation base
CPQ Citation Position Quality	6	14.6%	12.1%	5.7 – 25.1%	Most stable sub-metric; mean position varies little when citations are present
ASC Authority Source Coverage	6	15.3%	14.3%	8.7 – 26.0%	Moderate variance; authority-domain set is comparatively stable across days

F.4 Consequences for PS Interpretation

The empirically demonstrated daily variance confirms the methodological characterisation of PS as a semi-stochastic indicator. Concrete operational consequences:

1. **Single-day values are not reliable.** Mean CV of 14–30% across most sub-metrics shows that day-based statements ("PS improved yesterday") are not methodologically robust.
2. **Weekly or monthly means as standard.** For strategic statements, sub-metrics are averaged across at least 7 days (preferably the 30-day monthly average from the monitoring-tool CSV rolling window CSV rolling window).
3. **Effect-size thresholds with safety margin.** An M3 evaluation defines minimum effect sizes that must exceed the measured daily variance (e.g., ΔCVR ≥ 10% at a median CVR variance of 38% — an effect close to the median variance, which requires additional pre-trend validation per §5.4).
4. **Sub-metric-specific confidence.** CPQ with median 12% CV is the most stable metric and supports quarterly trend statements; CVR with median 38% CV supports only monthly or quarterly means, not weekly values.

F.5 Limitations

- **Sample size:** n = 6 clients is small. Cross-industry generalisability is limited; the study is updated with each additional pre-snapshot.
- **Time window:** 13 days is the lower bound for CV computation. A clean ICC study over 30+ days is on the v4.0 roadmap, planned for Q4/2026 once n ≥ 15 longitudinal cases are available.
- **Sub-metrics with constant zero value** (e.g., when a client appears in none of the days in CORE prompts) are excluded from CV computation because CV is undefined at mean 0. These clients structurally show high constancy at a low level, however.
- **Bias possibility:** The sample contains clients at different maturity levels (RS-Pre 33 to 88). Stratification by RS class is statistically meaningful only with n ≥ 10 per class.

This study addresses Major Revision MR3 from the external methodological review (Perplexity, 11 May 2026) as a first empirical step. A complete ICC study (inter-rater reliability with a longer time window and stratified sample) is the primary objective of the Impact Library build-up phase for v4.0.

COLOPHON

About this Document

Title	GEO-Score Framework v3.3.5 · Whitepaper (English Edition)
Version	3.3.5
Date	11 May 2026
Publisher	Johannes Bopp GmbH, Braunschweig, Germany
Operating brand	kmugeo · The GEO agency for the DACH industrial Mittelstand
Author	Tobias Ackermann (Head of GEO Operations & AI Visibility)
License	Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
Standard citation	Ackermann, T. (2026). GEO-Score Framework v3.3.5. Johannes Bopp GmbH. Available at kmugeo.de/geo-score-framework-en
German edition	Available at kmugeo.de/geo-score-framework

Citation block (NEW v3.3.1)

APA 7:

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

BibTeX:

```
@techreport{ackermann2026geo_en,
  author      = {Ackermann, Tobias},
  title       = {{GEO-Score Framework v3.3.5: Methodological Whitepaper (English
Edition)}},
  year        = {2026},
  month       = {May},
  institution = {Johannes Bopp GmbH},
  url         = {https://kmugeo.de/geo-score-framework-en},
  publisher   = {Zenodo},
  doi         = {10.5281/zenodo.20137223},
  url         = {https://doi.org/10.5281/zenodo.20137223},
  note        = {Creative Commons Attribution-ShareAlike 4.0 International}
}
```

Published via Zenodo on 12 May 2026 under DOI `10.5281/zenodo.20137223`. Reproducibility bundle (script + anonymised data) as ZIP within the same dataset.

Current Status (v3.3.5) — publishing-ready

Final-review response after the second external methodology review (Perplexity, 12 May 2026, FINAL DECISION: ACCEPT). Four targeted refinements: (1) inline reference to Appendix F in the PS-tracking chapter with daily-variance quantification; (2) new PS sub-metric stability matrix in Appendix A.0.b with CV values from Appendix F; (3) quantification note in Executive Summary Key Claim 2; (4) roadmap refinement ICC study Q4/2026 once n ≥ 15 longitudinal cases in Appendix F.5. Reviewer quote: With Appendix F, the framework is fully peer-review-ready for practice-scientific publications. Methodology unchanged.

Version History (compact)

VERSION	DATE	REASON
v3.3.5	12 May 2026	Final-review response (Perplexity ACCEPT): 4 targeted refinements + PS stability matrix
v3.3.4	12 May 2026	Consistency patch: tool-agnosticism in Appendix F + path reference
v3.3.3	12 May 2026	Empirical patch: Appendix F PS Daily-Variation Study (n=6, median CV 3.5–38.2%)
v3.3.2	12 May 2026	Privacy patch (DE anonymisation) + case extension 2→7 + pre-registration convention
v3.3.1	11 May 2026	Quick wins after Perplexity methodology review (5 minor revisions)
v3.3	11 May 2026	Roll-up bump CW 19/20: tool-agnosticism, B3.9 path separation, pre-trend test, pool stability filter
v3.2.3	9 May 2026	Hybrid model for DiD data source (client-individual pool §9.5 + industry whitelist)
v3.2.2	8 May 2026	Operational measurement mechanics (T+14/T+30 retrospective from 30-day rolling)
v3.2.1	8 May 2026	Patch release after three methodology review rounds (concept roles, blanket-caveat formula)

Full detailed descriptions of all bumps and the maintenance convention for future versions in the `CHANGELOG.md` co-published within this dataset.

Johannes Bopp GmbH · operating brand: kmugeo