A Reproducible Methodology for Generative Engine Optimization Measurement

Abstract

Sighted measures generative-engine visibility as an observed system, not a ranking proxy. The platform persists raw engine responses, extracted measurements, diagnostic events, and calibration metadata so customer-facing scores can be recomputed from source artifacts rather than inferred from dashboards after the fact.

This whitepaper documents the production methodology behind GEO Index, effectiveness validation, replica variance, ghost-citation detection, and fact-density scoring as of methodology version 1.0.0.

Engine coverage and selection criteria

The public GEO Index only scores engines with live adapters and stable evidence capture. The current eligible engine set is defined by the active row in geo_index_calibration_versions and exposed through /api/v1/index/methodology/current.

As of geo-index-v1.1, the eligible engine set is chatgpt, claude, perplexity, gemini, and google-ai-overviews. Additional engines may be visible elsewhere in the product, but they do not affect the GEO Index until calibration is activated for them.

Query panel construction and locked panel versions

Customer prompts are versioned through query panels rather than edited in place. GEO records the panel version used for each run so historical scores remain interpretable even if the panel evolves later.

That means methodological reproducibility requires both the engine response and the panel version that produced it. GEO stores the panel/query lineage through query_panels, queries, and run-linked metadata.

Replica design and sampling policy

Generative engines are non-deterministic. GEO therefore samples repeated replicas per run and treats replica count as part of the measurement design. Run-level outcomes are aggregated across replicas before intervals are computed across runs.

Replica-level artifacts include brand mentions, cited sources, retrieved sources, and diagnosis results. These are persisted through run-linked evidence tables and reused by diagnostics, GEO Index, crawler analytics correlation, and ghost-citation detection.

Scoring formulas (visibility, paper-faithful Princeton, GEO Index)

The deployed GEO Index implementation in packages/index/src/geo_index/service.py is a weighted mean of engine-level scores. The current component weights are:

- `share_of_voice = 0.30`
- `recommendation_rate = 0.20`
- `position_weighted_visibility = 0.20`
- `citation_rate = 0.15`
- `faithfulness = 0.15`

Each component is normalized against the active calibration version before engine-level composition. The customer-facing score is the mean of eligible engine scores, with intervals produced from bootstrapped samples of run-level composites.

RAID rewrite methods are kept separate from scoring so content interventions can be planned and measured without redefining the measurement contract after deployment.

Diagnostic rubric (discovery, retrieval, influence, faithfulness)

The primary diagnosis rubric has four stages:

- `discovery`: did the brand appear at all?
- `retrieval`: was the brand or its evidence retrieved when relevant?
- `influence`: did the brand move from mention to recommendation?
- `faithfulness`: was the brand represented accurately?

These stages power dashboard breakdowns, ghost-citation analysis, and action routing into RAID. The implementation anchors live in packages/monitor/src/geo_monitor/services/diagnosis.py and packages/monitor/src/geo_monitor/services/measurement_read.py.

Statistical treatment (bootstrap CIs, Wilson intervals, p-values, real-change threshold)

GEO uses bootstrap intervals for mean-like statistics and Wilson intervals for proportions. The shared implementation lives in packages/monitor/src/geo_monitor/services/statistics.py.

- Bootstrap mean and delta estimates use 1,000 resamples with replacement and a fixed seed for reproducibility in the public implementation.
- Wilson intervals are used for bounded proportions such as recommendation rate and citation rate because they remain better behaved than normal approximations at smaller sample sizes.
- Public p-values in effectiveness runs are two-sided bootstrap sign estimates.
- A real-change flag is only set when the observed mean delta is at least as large as the interval half-width.

Calibration process

Calibration observations are recorded in geo_index_calibration_observations over a rolling 90-day window. Those observations are then summarized into geo_index_calibration_versions, which defines the active engine set, normalization factors, and sample size.

Customer-facing scores in geo_index_scores reference the calibration version used at scoring time. New rows also store the public methodology version so documentation and persisted measurements stay aligned.

Known biases and limitations

- GEO Index is only as complete as the eligible engine set for the active calibration version.
- Public audit reproducibility depends on stored response artifacts and extracted measurements; engines that suppress citations constrain downstream analysis.
- Industry benchmarks cited in ghost-citation and fact-density documentation are external benchmark data, not internal causal claims.
- Replica averaging reduces variance; it does not eliminate model drift, retrieval shifts, or vendor-side ranking changes.

Listicle engine methodology

WS13 introduces a dedicated Listicle Generation Engine for Top-N act-surfaces. Its ranking weights, freshness rotation, gap-finder heuristic, and customer-position honesty contract are documented separately in [docs/methodology/listicle-engine.md](../methodology/listicle-engine.md).

Structured Data Stack

WS13 adds a structured-data control layer that treats schema as a measurable publication surface rather than a one-time markup task. The schema package generates JSON-LD only from extracted page facts, validates payloads against a vendored schema.org vocabulary plus Google rich-result rules, and blocks deployment whenever the validator returns an error-level issue. This means the platform prefers shipping no schema over shipping malformed schema that could create rich-result penalties.

The Triple Stacker coordinates Article, ItemList, and FAQPage generation for ranking-style pages, but it does so by composing the existing per-type eligibility rules rather than by inventing a second decision layer. Default behavior allows partial stacks when only some types are valid, while strict mode can require all three. Speakable extraction follows the same deterministic philosophy: it selects short declarative snippets from the primary article region and the highest fact-density paragraph, then limits the total selection to a short voice-ready budget.

WS13 also adds a site-wide scanner and a schema deployment path. The scanner compares live on-page JSON-LD against generated proposals, classifies findings using existing eligibility and validator outputs, and stores stable diffs so repeated unchanged scans remain byte-identical. Deployment routes through the existing integrations and approvals stack, which means schema changes inherit the same rollback and governance guarantees as other content publication actions.

Schema experimentation intentionally reuses the existing effectiveness framework instead of defining a separate statistics engine. Schema experiments stamp cohort metadata into Run.model_config_snapshot, exclude ambiguous cross-variant runs from membership, and read the existing GEO_EFFECTIVENESS_SIGNIFICANCE_LEVEL threshold when deciding winners. Before an experiment can start, WS13 runs a pre-flight power check based on the workspace's recent eligible run history; underpowered workspaces are refused up front rather than running a calendar window with no realistic chance of learning.

Finally, WS13 extends discoverability artifacts in place. llms.txt v2 ranks pages with a blended score over demand, citation volume, freshness, and operator intent; llms-full.txt adds deterministic excerpts under the same priority ordering; and robots.txt plus bot directives are rendered deterministically from the crawler taxonomy so artifact history reflects real policy changes instead of formatting noise. The underlying methodology details are documented separately in [docs/methodology/schema.md](../methodology/schema.md) and [docs/methodology/discoverability.md](../methodology/discoverability.md).

Reproducibility bundle specification

Each completed public audit exposes a reproducibility bundle at /api/v1/public/audit/{request_id}/reproducibility-bundle. The bundle contains:

- `methodology_version`

- `audit_id`
- `raw_engine_responses`
- `computed_scores`
- `formula_references`

The bundle can be passed back through the public-audit recomputation helper to confirm that the published score matches the persisted calculation.

References

1. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). *GEO: Generative Engine Optimization*. KDD 2024. DOI: https://doi.org/10.1145/3637528.3671900 2. Seer Interactive. (2026-03-24). *LLM Ghost Citations: Why Your Content Is Working and Your Brand Isn't*. https://www.seerinteractive.com/insights/llm-ghost-citations-why-your-content-is-working-and-your-brand-isnt 3. Erlin. (2026). *AI Brand Visibility Tracking: A Complete Guide (2026)*. https://www.erlin.ai/blog/ai-brand-visibility-tracking