SA-AEO-Bench v1 — OSF Pre-Registration Protocol

Working title: SA-AEO-Bench v1: An Open Benchmark for Measuring LLM Citation Behavior on South African Queries

Research entity: Cited Brands (citedbrands.co.za) — wannabdad@gmail.com **Author / Principal investigator:** Joseph K Banda, Cited Brands **Affiliation:** Cited Brands — independent research initiative, South Africa **OSF pre-registration:** https://osf.io/w4az2 (public, immediate, no embargo) **Pre-registration date:** 2026-05-19 **Data collection start:** On or after 2026-05-19 (post-registration) **Data collection end:** Expected within 14 days of start **Public report URL:** citedbrands.co.za/research/sa-aeo-bench-v1 (TBD)

1. Study purpose

To produce the first publicly registered, reproducible map of how three frontier Large Language Models (LLMs) — OpenAI GPT-5, Anthropic Claude Sonnet 4.5, Google Gemini 2.5 Pro — cite source material when answering South Africa–specific queries. The benchmark and dataset will be released under CC-BY-4.0 to seed a reusable evaluation harness for the emerging Answer Engine Optimization (AEO) field.

Prior published evidence on LLM citation behavior is concentrated on US-centric queries (Liu et al., 2023; Stanford HELM, 2024). No published study reports citation behavior on Southern African queries. This protocol fills that gap.

2. Pre-registered hypotheses

These hypotheses were specified before data collection and will be tested using the analysis plan in §6. We commit to reporting all results regardless of statistical significance.

ID	Hypothesis	Predicted direction
H1	SA-domain citation share will exceed 50% for organic (blind) SA category queries, averaged across the 3 LLMs	one-sided, > 50%
H2	The top-20 cited domains will differ significantly between the 3 LLMs	two-sided (chi-square of independence)
H3	Latin Square position-swap on comparison questions will produce mean Jaccard citation-overlap < 0.6 across the 3 LLMs (i.e. >40% of cited domains will change when prompt order is reversed)	one-sided, < 0.6
H4	For brand-authority questions, SA-domain citation share will be lower for *negative* polarity prompts than for *positive* polarity prompts, by ≥10 percentage points	one-sided
H5	Non-English SA queries (Afrikaans, isiZulu) will return citations of which ≥80% point to English-language content	one-sided, ≥ 80%
H6	Reddit citation count by Gemini will exceed Reddit citation count by GPT-5 and by Claude, each, by a factor of ≥5×	one-sided, ratio ≥ 5
H7	Cited-domain Jaccard overlap between capped and uncapped Gemini responses on the same prompts ≥ 0.70 (i.e. the search-budget standardization in §3.5.1 does not materially distort the citation map)	one-sided, ≥ 0.70

Each hypothesis includes a directional prediction. Results will be reported as confirmed / disconfirmed / inconclusive against the stated direction.

3. Sample design

3.1 Industries

Ten SA consumer industries, selected for breadth across financial, telecom, retail, automotive, and lifestyle categories:

#	Industry	Brands sampled (n=10 per industry)
1	Banking	FNB, Standard Bank, Capitec, Absa, Nedbank, Investec, African Bank, TymeBank, Discovery Bank, Bidvest Bank
2	Telecom / mobile	Vodacom, MTN, Cell C, Telkom, Rain, Afrihost, Webafrica, MWeb, RSAWeb, Vox
3	Grocery retail	Pick n Pay, Checkers, Woolworths Food, Spar, Shoprite, Food Lover, Makro, Boxer, Game Food, OK Foods
4	Medical aid	Discovery Health, Bonitas, Momentum Health, Medihelp, Fedhealth, Profmed, GEMS, Bestmed, Sizwe Hosmed, KeyHealth
5	Short-term insurance	OUTsurance, MiWay, Hollard, Naked Insurance, Santam, Dialdirect, Discovery Insure, King Price, Budget Insurance, 1st for Women
6	Automotive / EV / dealers	Toyota SA, VW SA, Ford SA, Hyundai SA, Suzuki SA, Mercedes-Benz SA, BMW SA, BYD SA, GWM ORA, Motus / AutoTrader
7	E-commerce	Takealot, Superbalist, Yuppiechef, Bash, Loot, Zando, Bidorbuy, OneDayOnly, Cape Union Mart, Game Online
8	Restaurants / food delivery	Mr D, Uber Eats SA, Nando’s SA, Steers, Spur, KFC SA, McDonald’s SA, Roman’s Pizza, Debonairs, Pedros
9	Streaming / entertainment	Showmax, DStv, Netflix SA, Amazon Prime Video SA, Apple TV+ SA, MultiChoice, Disney+ SA, eVOD, BoxOffice, Vidi
10	Real estate / property	Pam Golding, Seeff, RE/MAX SA, Property24, Private Property, ooba Home Loans, Just Property, Tyson Properties, Engel & Völkers SA, Lew Geffen Sotheby’s

Total: **10 industries × 10 brands = 100 brands**.

3.2 Question structure

Each brand receives **10 prompts** spanning 4 question types, designed to cover the full buyer-journey intent space:

#	Type	Polarity	Prompt template
Q1	Organic blind	n/a	“What’s the best {industry} in South Africa in 2026? Give specific recommendations with source URLs.”
Q2	Organic scenario	n/a	“My friend is looking for a {industry} in South Africa. What would you recommend and why?”
Q3	Brand authority	positive	“Is {brand} actually any good as a South African {industry}? I want honest opinions with sources.”
Q4	Brand authority	negative	“What are the biggest complaints about {brand} as a South African {industry}? I want to know the downsides — cite sources.”
Q5	Brand authority	balanced	“Give me an honest pros-and-cons assessment of {brand} as a South African {industry}, citing sources.”
Q6	Comparison	Latin Square A vs B (competitor 1)	“I’m choosing between {brand} and {competitor_1} for {industry} services in South Africa. Which should I pick and why? Cite sources.”
Q7	Comparison	Latin Square B vs A (competitor 1)	“I’m choosing between {competitor_1} and {brand} for {industry} services in South Africa. Which should I pick and why? Cite sources.”
Q8	Comparison	Latin Square A vs B (competitor 2)	Same template, with {competitor_2}.
Q9	Comparison	Latin Square B vs A (competitor 2)	Reversed order, with {competitor_2}.
Q10	Value / pricing	n/a	“Is {brand} worth the cost as a South African {industry} in 2026? Cite sources comparing value vs competitors.”

`{competitor_1}` and `{competitor_2}` = next two brands in the industry list, wrapping at the end.

3.3 Replications

Every prompt is executed **5 times** (sent to each of 3 models on 5 separate API calls per model). This enables:

1. **Bootstrap confidence intervals** on every per-prompt metric (citation count, SA-share, top-domain hit rate)
2. A **citation-stability metric** — % of cited domains that recur across the 5 replications for the same prompt + model. This is itself a pre-registered finding: we expect Gemini’s stability < GPT-5’s < Claude’s, reflecting search vs training knowledge.
3. Robustness against LLM stochasticity at non-zero temperature

3.4 Multilingual subsample

100 organic queries (50 Afrikaans + 50 isiZulu) distributed across the 10 industries (5 per language per industry). Each is replicated 5 times. Total multilingual = **500 API-calls-per-model**.

3.5 Total sample size

- Base English prompts: 10 industries × 10 brands × 10 questions = **1,000 unique prompts**
- × 5 replications each = **5,000 English prompts**
- Multilingual: 100 unique × 5 replications = **500 multilingual prompts**
- **Grand total: 5,500 prompts × 3 models = 16,500 API calls**

3.5 Models

Model	Provider	Web retrieval mechanism	Why selected
GPT-5-chat	Azure OpenAI	`web_search` tool	Largest deployed user base on ChatGPT
Claude Sonnet 4.5	Azure Anthropic Foundry	`web_search_20250305` tool, max_uses=2	Second-largest frontier model by usage
Gemini 2.5 Pro	Google AI Studio (direct)	`google_search` grounding	Google AI Overviews backbone — highest user reach via Search

Each prompt is sent to all 3 models. Models receive identical prompt text — no prompt customization per model.

3.5.1 Cross-provider search-budget standardization

To enable fair per-search-cost comparison across providers and limit unbounded grounding costs, search budget is standardized to ~2 web searches per call:

- **Claude:** API parameter `max_uses: 2` on the web_search tool
- **GPT-5:** no explicit cap (averages ~1 search/call by default; no clamp applied)
- **Gemini:** appended instruction “Use at most 2 web searches to answer.” Validated in a 60-call pre-study (2026-05-19) to reduce mean searches from 7.53 to 2.57 without material reduction in grounding chunk count (–9%) or response length (–8%).

A pre-registered 200-call uncapped Gemini subsample will validate that the cap does not materially distort the citation map (H7).

H7 | Cited-domain overlap (Jaccard) between capped and uncapped Gemini responses on the same prompts ≥ 0.70 | one-sided, ≥ 0.70 |

3.5.2 Foreknowledge of data and pilot disclosure

A 115-prompt methodology pilot was conducted on 19 May 2026 prior to this pre-registration. The pilot's purpose was strictly engineering and methodology validation:

- Confirming API integration with all three providers (Azure OpenAI, Azure Anthropic Foundry, Google Gemini)
- Validating citation-URL extraction code for each provider's distinct response shape
- Refining the 12-category URL classification rubric in §7
- Verifying that the Gemini search-budget cap (§3.5.1) reduces searches without materially distorting citation quality

The pilot dataset is NOT part of the 16,500-call analysis dataset specified in this registration and will not enter any H1–H7 confirmatory analysis.

Two methodological choices in this registration were informed by pilot findings:

1. The Gemini search-budget cap was validated by a separate 60-call A/B test (capped vs uncapped) before adoption. This is hypothesis H7 in the current registration, which will be re-tested on a fresh 200-prompt subsample drawn from the 16,500-call main dataset.
2. The URL classification rubric's 12 categories were initially specified based on observed pilot citations, then frozen in this pre-registration before the main study begins. The frozen rubric, with full domain inclusion lists, is committed to the public repository at the time of OSF registration.

Risk mitigations beyond pre-registration:

- Protocol, prompts, code, and rubric are all open-source and timestamped
- Inter-rater reliability check on URL classification (Cohen's kappa target ≥ 0.80, §6.4) detects rubric ambiguity post hoc
- Pre-declared exclusion criteria (§5) prevent ad-hoc filtering of the main dataset

3.5.3 Study type classification

This is a **non-randomized observational study** combined with **descriptive measurement**. Independent variables (LLM, industry, brand, question type, language, comparison order) are systematically varied across conditions in a factorial design but are not randomly assigned in the experimental-control sense. Subjects are LLM API endpoints, not human participants.

Intention for causal interpretation: None. This study makes no causal claims. We measure WHAT each LLM cites for SA queries and describe the distributions, differences, and stability of those citation patterns. We do not infer that any property of an LLM causes its citation behavior.

3.5.4 Randomization and counterbalancing

No random assignment of subjects to treatments. In place of randomization, two systematic counterbalancing controls:

1. **Latin Square counterbalancing** on all comparison questions (Section 3.2, hypothesis H3). Every "brand A vs brand B" prompt is asked in both A-B and B-A orderings, making position-

of-mention bias a measured quantity rather than a confounder.
2. **Five replications per unique prompt** (Section 3.3). Each (prompt × model) pair is fired 5 times, generating bootstrap-resampleable variance estimates and a citation-stability score per model. Replication mitigates LLM stochasticity at non-zero temperature without requiring randomized assignment.

Brand-to-competitor pairings use deterministic next-brand rotation within each industry (with wrap-around at the end of the list). The prompt sequence executed by the runner script is deterministic given the protocol and is reproducible from the committed code.

3.5.5 Blinding mechanisms

Two blinding mechanisms in active use during data collection and analysis:

1. **Blind prompts.** For organic questions (Q1, Q2 in §3.2), the target brand is never named in the prompt text. The LLM responds to a category-level question with no information about which brand is the focal subject of that data row. This measures unprompted brand visibility — the LLM cannot be cued to favor a specific brand because the brand name is not in the input.

2. **Blind URL classification.** A 5% random subsample of cited URLs is re-classified by an independent classifier (a second LLM acting as classifier). The classifier receives only the URL string itself, blinded to which originating LLM cited the URL. Inter-rater reliability is reported as Cohen's kappa with a target of ≥ 0.80. This eliminates the risk that the 12-category URL classification rubric is applied differently for URLs from different LLMs.

Latin Square counterbalancing (§3.5.4) is also a form of blinding-equivalent control: position-of-mention bias becomes a measured quantity (H3) rather than a confounder.

3.6 Data collection window

All API calls will be completed within a 14-day window. Model version IDs and Azure deployment names will be logged with every response.

4. Procedure

1. Prompts are generated programmatically from the brand/industry/template specification before data collection begins. The full prompt set is published to GitHub at OSF pre-registration time (committed before any API calls).
2. Each prompt is sent exactly once to each model. **No batching** — one question per API call to preserve citation attribution and independence assumptions.
3. Web search / grounding is enabled on every call.
4. Responses are captured with:
    - Full response text
    - All citation URLs and titles
    - Provider's reported token counts
    - Number of web searches performed
    - Wall-clock latency

    - Any error codes / partial responses
5. Vertex AI redirect URLs returned by Gemini are resolved via HEAD request to the final destination domain before classification.
6. URLs are classified into 12 categories per the rubric in §7.

5. Inclusion / exclusion criteria

Included: all successful API responses (HTTP 200) with a non-empty body.

Excluded from primary analysis (reported separately as data-quality metrics): - API calls with HTTP 4xx or 5xx errors after 3 retry attempts - Responses with zero citation URLs returned (counted separately as “no-citation rate”) - URLs that fail to parse into a valid hostname - Vertex redirect URLs that fail to resolve after 5 seconds

6. Analysis plan

6.1 Primary analyses (linked to hypotheses)

Hypothesis	Test	Inference threshold
H1	Bootstrap 95% confidence interval (BCa method, 1,000 resamples) on SA-domain citation share	Confirmed if lower bound > 50%
H2	Chi-square test of independence on top-20 domain × model contingency table	α = 0.01 (Bonferroni-corrected)
H3	Mean Jaccard citation-overlap across all Latin Square pairs, with bootstrap 95% CI	Confirmed if upper bound < 0.6
H4	Per-brand SA-share difference (negative – positive); paired bootstrap 95% CI	Confirmed if upper bound < –10pp
H5	% of multilingual-query citations resolving to English-language pages (manual	Confirmed if lower bound ≥ 80%

Hypothesis	Test	Inference threshold
	review of cited page language on 100% of multilingual citations)	
H6	Ratio of Reddit citations: Gemini / max(GPT-5, Claude)	Confirmed if ratio ≥ 5
H7	Jaccard of cited domains: 200-prompt capped Gemini subsample vs same prompts run uncapped, paired bootstrap 95% CI	Confirmed if lower bound ≥ 0.70

6.2 Secondary analyses (exploratory, declared in advance)

- Per-industry SA-share with 95% CI
- Per-industry top-10 cited domains
- Per-brand visibility rate in organic queries (Bradley-Terry MLE strength scores)
- Sycophancy uplift per brand (named-probe SA-share vs blind-probe SA-share for same brand)
- Per-model "exclusive" domains (cited by exactly one of the 3 models)
- Per-model "consensus" domains (cited by all 3 models)
- Latin Square verbal-winner flip rate (qualitative; whether the brand named first in the prompt is the brand named first in the response)
- **Citation-stability score per model** — for each (prompt × model) pair, compute the Jaccard overlap of cited domains across the 5 replications, then average. Higher score = more stable / deterministic retrieval. Pre-declared expected order: Claude > GPT-5 > Gemini (Gemini's broader search budget produces more diverse runs)

6.3 Replication-based confidence intervals

All point estimates in the primary analysis will be accompanied by bootstrap 95% confidence intervals computed at the *prompt* level: each of the 1,100 unique prompts contributes 5 replicated

observations per model. Bootstrap resamples are taken at the prompt level (not the call level) to avoid pseudo-replication.

6.3 Inter-rater reliability for URL classification

A 5% random subsample of unique URLs (~250 URLs from the expected ~5,000 unique domains) will be classified by a second independent classifier — either a different LLM (Gemini 2.5 Pro acting as classifier on URLs cited by other models, blinded to which model cited each URL) or a human rater.

Threshold: Cohen's kappa ≥ 0.80 on the 12-category rubric. If kappa < 0.80, the rubric will be revised, re-applied to the full dataset, and the revision noted in the report.

7. URL classification rubric (12 categories)

Category	Rule
`sa_local_press`	Domain in curated SA-publisher list (news24, iol, dailymaverick, timeslive, businesslive, fin24, moneyweb, mybroadband, bizcommunity, citizen, businesstech, dailyinvestor, etc. — full list committed to GitHub)
`sa_review`	hellopeter.com
`sa_directory`	brabys.com, snupit.co.za, whocandoit.co.za
`sa_gov`	hostname ends with `.gov.za`
`sa_edu`	hostname ends with `.ac.za` or `.edu.za`
`sa_other`	hostname ends with `.co.za`, `.org.za`, `.net.za`, or `.web.za` and not in above categories
`wikipedia`	wikipedia.org or any language subdomain
`reddit`	reddit.com
`youtube`	youtube.com or youtu.be
`social`	facebook.com, linkedin.com, twitter.com, x.com, instagram.com
`intl_review`	trustpilot.com, yelp.com, tripadvisor.com, g2.com, capterra.com, complaintsboard.com, sitejabber.com
`intl_press`	Curated international publisher list (nytimes, reuters, bloomberg, ft, wsj, cnn, bbc, theguardian, forbes — full list on GitHub)
`intl_other`	Anything not matching above

Categories are **mutually exclusive**. Classification precedence: exact-match lists first, then SA-TLD pattern, then international fallback.

“SA citation” in any aggregate = any category with prefix `sa_`.

8. Reproducibility commitments

1. **Code:** all data collection, classification, analysis, and reporting scripts published to a public GitHub repository (github.com/citedbrands/sa-aeo-bench) at the time of OSF pre-registration submission. Repo URL committed in OSF registration.
2. **Prompts:** the full set of 5,500 prompts (1,100 unique × 5 replications) is generated by a deterministic script and committed to the repo before any API call is made. Anyone can re-generate the exact prompt set from the script.
3. **Dataset — released:** Aggregate citation dataset published under CC-BY-4.0 at report publication, including (per model × industry × question type × language):
    - Per-domain citation counts
    - SA-share metrics with 95% bootstrap CIs
    - Top-cited domain rankings
    - Latin Square Jaccard distributions
    - Citation-stability scores across replications
4. **Dataset — retained:** Raw response text (the verbatim LLM output for each of the 16,500 API calls) is retained by Cited Brands for use in subscriber-facing commercial products (per-brand audits, recommended-publication-target lists). This split is fully disclosed in §11 (Conflict of Interest). Aggregate-only data release is standard practice in industry-funded research where raw outputs contain commercially sensitive material; the methodology + code + aggregate data are sufficient for any third party to independently re-run the study and verify findings.
5. **Model versions:** exact deployment IDs, version strings, and API endpoint URLs logged in every response record. Reported in the methodology section of the public report.
6. **Run window:** start and end timestamps published, so the dataset reflects a known snapshot in time.
7. **Re-runnability:** Because all code, prompts, and model identifiers are public, any researcher with API access to GPT-5, Claude Sonnet 4.5, and Gemini 2.5 Pro can re-execute the study (cost: ~$1,450) and compare findings.

9. Pre-declared exclusions from the conclusions

- Findings about specific brands' competitive positioning relative to each other beyond what the BT-Elo analysis shows. We will not publish per-brand "rankings" outside the BT-Elo framework.
- Causal claims about why models cite specific sources. The benchmark measures *what* they cite, not why.

10. Timeline

Milestone	Date
OSF pre-registration submitted	T+0
Public GitHub repo created with prompts + code	T+0
Data collection start	T+7 (minimum 7-day delay from pre-registration to ensure protocol can be reviewed)
Data collection end	T+21
Inter-rater classification pass	T+25
Analysis complete	T+30
Report draft	T+35
Report published	T+42

11. Funding and conflict of interest

Cited Brands (citedbrands.co.za) is funding the API costs (~R26,500 / ~$1,450 USD) directly. Cited Brands operates research-driven commercial products in the SA AEO measurement space; this benchmark will be used both as commercial evidence and as a public reusable artifact. The conflict is disclosed upfront. To mitigate: protocol, prompts, code, classification rubric, and aggregate dataset are released openly (per §8), allowing any third party to re-run, verify, or extend the study at an independent cost of ~$1,450.

12. Authors' commitment

We commit to publishing the full report — including all tests of pre-registered hypotheses — within 90 days of data collection completion, regardless of whether results favor or disfavor any commercial interest of GenPicked. If any analysis is added after data inspection, it will be clearly labeled as exploratory.

This protocol follows the structure of the Pre-Registration of Quantitative Studies (PRP-Quant) template (Bosnjak et al., 2022). The final OSF registration will use OSF's standard form, mapping each section above to the corresponding OSF field.

References

- Liu, N.F. et al. (2023). "Lost in the Middle: How Language Models Use Long Contexts." *Transactions of the Association for Computational Linguistics*. arXiv:2307.03172.

- Stanford CRFM (2024). “Holistic Evaluation of Language Models (HELM).” crfm.stanford.edu/helm
- LMSYS (2024). “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference.” arXiv:2403.04132.
- Bosnjak, M., et al. (2022). “A template for preregistration of quantitative research in psychology.” *Advances in Methods and Practices in Psychological Science*.
- Fisher, R.A. (1935). *The Design of Experiments*. Oliver and Boyd. (Latin Square methodology)
- Bradley, R.A., Terry, M.E. (1952). “Rank Analysis of Incomplete Block Designs.” *Biometrika*.