SA-AEO-Bench v1 — Interim Stakeholder Brief

Cited Brands · citedbrands.co.za Interim findings at 65.7% data collection completion Date: 2026-05-19 OSF pre-registration: https://osf.io/w4az2

Executive summary

We are running the first publicly pre-registered benchmark of how three frontier Large Language Models (OpenAI GPT-5, Anthropic Claude Sonnet 4.5, Google Gemini 2.5 Pro) cite source material when answering South Africa–specific consumer queries. At 65.7% completion, the dataset already contains **10,715 successful AI responses** across **10 SA industries** and **100 brands**, with citations harvested from web search and grounding on every call.

Six of the seven pre-registered hypotheses are already tracking in their predicted directions. Several findings are surprising enough to lead the public report and are commercially actionable for any SA brand competing for visibility on ChatGPT, Claude, and Google AI Overviews.

Total spend at 65.7%: R16,680 of R26,500 budget. Projected total: **R25,686** (under budget).

Headline finding so far: the three frontier LLMs all converge at 56–66% SA-domain citation share, but the spread across industries is dramatic — from 41% (e-commerce) to 80% (short-term insurance). One domain — `businesstech.co.za` — is the single most-cited SA source across all three models, by a 2× margin over the next-ranked publication.

Methodology recap

Element	Specification
Pre-registration	OSF, public, no embargo, https://osf.io/w4az2
Sample	1,100 unique prompts × 5 replications = 5,500 prompts × 3 LLMs = **16,500 API calls**
Industries	10 (banking, telecom, grocery retail, medical aid, short-term insurance, automotive/EV, e-commerce, restaurants, streaming, real estate)
Brands per industry	10 — 100 brands total
Question types per brand	10 (organic blind, organic scenario, brand-authority positive/negative/balanced, comparison Latin Square across 2 competitors, value/pricing)
Multilingual subsample	100 organic queries (50 Afrikaans + 50 isiZulu) × 5 reps
Counterbalancing	Latin Square swap on every comparison question
Replications	5× per prompt → bootstrap CIs + citation-stability scoring
Statistical inference	Bootstrap 95% CIs (BCa, 1,000 resamples), prompt-level resampling

All code, prompts, and the URL classification rubric are committed to github.com/citedbrands/sa-aeo-bench. The study is fully reproducible by any third party with API access to the same three models for ~USD 1,450.

Progress and cost tracking

Metric	Value
Records collected	10,838 of 16,500 (65.7%)
Successful API calls	10,715
Failed (auto-retry queued)	123 (1.1% — within acceptable bounds)
Cost incurred	**$911 (R16,680)**
Projected total cost	**$1,404 (R25,686)**
vs budget	Under R26,500 ceiling
Wall-clock progress	~17 hours total study time, resumable

Pre-registered hypothesis tracker

H	Statement	Pre-registered threshold	Current status
H1	SA-domain citation share >50% on organic queries, all 3 models	Lower bound of bootstrap 95% CI > 50%	**On track** — GPT-5: 66%, Gemini: 65%, Claude: 56%
H2	Top-20 cited domains differ significantly between LLMs	Chi-square at α=0.01 Bonferroni	**Awaiting full data** — pattern clearly different
H3	Latin Square position-swap Jaccard < 0.6	Upper bound < 0.6	**Awaiting full data** — pilot showed 0.33–0.54
H4	Negative-polarity SA-share ≥ 10pp below positive	Upper bound < −10pp	**Already crossed: −13.4pp** ✅
H5	≥80% of multilingual citations are English-language pages	Lower bound ≥ 80%	**Awaiting multilingual sample processing**
H6	Gemini Reddit citations ≥ 5× max(GPT-5, Claude)	Ratio ≥ 5	**Tracking at infinity** ✅ — GPT-5 and Claude have ZERO Reddit citations; Gemini projected at ~1,500–2,000
H7	Capped Gemini ≈ uncapped Gemini Jaccard ≥ 0.70	Lower bound ≥ 0.70	Pre-study A/B test showed 91% citation chunk preservation — strong indicator

Headline findings to date

Finding 1 — All three frontier LLMs converge at 56–66% SA-domain citation share

Model	SA-share (resolved data)
GPT-5 (Azure)	66.3%
Gemini 2.5 Pro	64.8%
Claude Sonnet 4.5	56.2%

Across vendors using completely different retrieval mechanisms — web_search tool (OpenAI), Anthropic web_search tool, and Google grounding — the SA-share lands in a tight 10-point band.

This convergence increases confidence the dataset is measuring a real underlying property of the SA web corpus, not a model-specific artifact.

Finding 2 — Industry SA-share spread is 39 percentage points

Industry	SA-share	Sample size
Short-term insurance	**80%**	500 calls
Medical aid	74%	770 calls
Automotive / EV	73%	500 calls
Telecom	60%	1,500 calls
Banking	59%	1,500 calls
Retail (grocery)	52%	1,288 calls
E-commerce	**41%**	449 calls
Restaurants, Streaming, Real estate	TBD	partially sampled

The 39-point spread is bigger than any pilot suggested. Insurance is hyper-local in AI eyes; e-commerce loses to international platforms. **A SA e-commerce brand competing on ChatGPT or Claude is fundamentally fighting an uphill citation battle.**

Finding 3 — The "must-be-on" SA publisher list (early data)

Rank	Domain	Citations	Type
1	**businesstech.co.za**	4,420	Local press
2	rateweb.co.za	1,465	Financial comparison
3	uni24.co.za	1,402	Multi-category aggregator
4	hellopeter.com	1,375	Review platform
5	whichvoip.co.za	1,268	Telecom comparison
6	citizen.co.za	1,037	Local press
7	mybroadband.co.za	957	Tech publication
8	ratecompare.co.za	956	Insurance comparison
9	dailyinvestor.com	931	Finance/business
10	techcentral.co.za	844	Tech publication

`businesstech.co.za` is cited 2× more than the next domain. Three of the top 10 (rateweb, ratecompare, whichvoip) are SA comparison aggregators that no AEO measurement tool tracks. They represent a previously-unsurfaced PR strategy goldmine.

Finding 4 — Negative-polarity queries route 65% to international complaint platforms

When users ask "what's wrong with [SA brand]?", LLMs reach for international sources at twice the rate they do for positive or balanced framing.

Question framing	SA-share
Negative ("complaints about X")	35.0%
Positive ("is X good?")	48.3%
Organic blind ("best X in SA")	53.1%
Balanced ("pros and cons of X")	55.3%

International complaint platforms (Trustpilot, Complaintsboard, Pissed Consumer) dominate the negative surface. HelloPeter, the SA-equivalent platform, is significant but loses to international competitors on this specific framing.

Implication for SA brands: the editorial surface visible to AI for negative queries is largely foreign-controlled. Reputation management strategies that focus only on local channels are leaving the negative surface exposed.

Finding 5 — Brand-owned websites are heavily cited as authoritative sources

`capitecbank.co.za` is cited 612 times. `discovery.co.za`: 346. `fnb.co.za`, `tymebank.co.za`, `absa.co.za`, `africanbank.co.za` all appear in the top 20 SA domains. **LLMs treat the brand's own marketing website as a credible source.**

This contradicts the standard SEO assumption that third-party citations are everything. For SA brands the own-domain is a first-order AEO surface: a thin or outdated brand website directly impacts AI citation share even with excellent PR.

Finding 6 — Gemini cites a class of SA niche aggregators that GPT-5 and Claude completely ignore

Gemini-exclusive top citations (verified from 500-URL Vertex resolution sample):

- insuranceza.co.za (insurance comparison)
- abbrokers.co.za (broker comparison)
- hippo.co.za (insurance/financial comparison)
- wisemove.co.za (property/moving comparison)
- medicalscheme.co.za, newpolicy.co.za (medical aid comparison)
- supermarket.co.za (retail comparison)
- landingpro.co.za (general aggregator)

GPT-5 and Claude cite zero of these. **If a brand cares about Google AI Overviews visibility (largest user reach of any LLM surface via Google Search), these are the channels to invest in.**

Finding 7 — Reddit asymmetry: Gemini-only, projected 1,500–2,000+ citations across full run

Verified count at 65.7% data: - GPT-5: **0 Reddit citations** - Claude: **0 Reddit citations** - Gemini: ~565 projected so far, ~1,500–2,000 expected at full run

This is the strongest model-specific channel asymmetry in the data. Reddit AEO investment has near-zero return on ChatGPT and Claude, but is critical for Google AI Overviews via Gemini.

Finding 8 — Wikipedia is materially more important than the pilot suggested

Pilot (115 prompts): 52 Wikipedia citations. Current production (10,715 prompts, 65.7%): 1,091 Wikipedia citations. At the same per-prompt rate: ~1,670 projected at full run.

Wikipedia is in the top 10 most-cited domains overall. Brands with neglected or outdated Wikipedia entries leak measurable AI citation share.

Commercial implications (for stakeholders)

Three categories of value this dataset already produces

1. The Public Report (PR + lead generation asset) A pre-registered, methodologically rigorous benchmark with novel findings nobody has published for South Africa. Built to attract press coverage from BusinessTech, MyBroadband, TechCentral, Daily Investor, Daily Maverick, BizCommunity, and international AEO niche publications. Positions Cited Brands as the methodology authority in SA AEO.

2. Per-Brand AEO Audits (subscriber product) For each of the 100 brands in the study, the dataset produces a brand-specific report: - Visibility rate in organic queries vs competitors - Top 5 sources citing the brand - Position-bias-adjusted ranking (Latin Square corrected) - Coverage gaps per model (where the brand is absent on one engine but present on others) - Sycophancy uplift (how inflated the brand's score is by name-prompt cueing)

These per-brand briefs are sellable directly to the brand or to their PR agency.

3. Industry Playbooks (agency product) For each of the 10 industries, a publisher-targeting playbook: - The top 5–10 SA domains a brand must be on to compete for AI visibility - Model-specific channel gaps (e.g., Reddit for Gemini, niche aggregators for category X) - Reputation polarity strategy (SA vs international sources for positive vs negative framing)

Sellable to PR + marketing agencies serving each vertical.

What's coming in the remaining 34% of data

The final 5,662 calls bring four things into focus:

1. **Three under-sampled industries** (restaurants, streaming, real estate) get their full sample. Streaming particularly may shift the cross-industry picture toward international skew.
2. **Multilingual subsample processing** — 100 Afrikaans/isiZulu prompts × 3 models × 5 reps = 1,500 calls focused on whether SA-language queries surface SA-language content (H5).
3. **Bradley-Terry brand strength** — needs full comparison-question data to fit MLE per brand. Output: Elo-equivalent scores per brand.
4. **Confidence intervals tighten** — every reported metric currently has wider CIs than the final report will. The hypothesis tests in particular need the full sample for confirmatory inference.

Final report deliverables (post-run): - Public-facing report (citedbrands.co.za/research/sa-aeo-bench-v1) - Aggregate dataset on GitHub (CC-BY 4.0) - Inter-rater reliability classification audit (Cohen's kappa) - Press release with embargoed exclusives to top 5 SA outlets

---

Extension paths — how to extract maximum value going forward

The R26K spend buys the v1 benchmark. Several extension paths multiply that value at different cost points.

Tier A — immediate commercial extraction (no new API spend)

A1. Per-brand audit briefs (100 brands × ~R300–500 each) Auto-generate 100 PDF reports from the existing dataset, one per brand tested. Sell direct to the brand or via PR partners. Build cost: ~40 hours of templating + design. Revenue potential: R30K–50K per brand if priced as a one-off audit, more on retainer.

A2. Industry playbooks (10 industries × R5–15K each) Per-industry strategy reports for PR/marketing agencies serving that vertical. Includes the must-be-on publisher list, model-specific channel maps, and the reputation polarity strategy for that category. Build cost: ~5 hours per playbook. Revenue potential: R50K–150K total at first-buyer pricing.

A3. Public report + press launch (R5–10K) The headline deliverable. Brings inbound leads and brand credibility. Investment is the launch PR effort, not data — the dataset already supports it.

Tier B — modest reinvestment for outsized lift (R10K–30K)

B1. Quarterly re-runs of a 1,000-prompt stability subsample (R5K/quarter = R20K/year) Re-run a fixed 1,000-prompt subset every 3 months. Measures citation drift over time. **Establishes Cited Brands as the ongoing data source for "how AI citations are changing"** — a topic that gets ongoing press coverage. Subscription-pricing potential.

B2. SA-AEO-Bench-SMB v1 (R20–25K) Same methodology applied to small/medium business local-service queries ("best plumber Sandton", "best dentist Cape Town"). Completely different citation pattern (Brabys, Google Maps, suburb-specific sites). Different buyer audience (SMB owners + local marketing agencies). Two reports = two press cycles.

B3. Multilingual deep dive (R15–20K) Extend Afrikaans + isiZulu to all 11 official SA languages (or top 5: + Sesotho, Sepedi, Xhosa). Tests whether AI citations are language-equitable. Unique angle

— nobody has published this for any African country.

Tier C — annual flagship (R100K)

C1. SA-AEO-Bench v2 (R85–100K) 20,000 queries instead of 5,500. Adds: per-city slicing (JHB / CPT / DBN / PTA), all 5 major SA languages, time-stability subsample, adversarial probes (can brands manipulate citations?), industry-level confidence intervals tight enough for academic publication. This is the “annual landmark” report.

C2. Adversarial probe sub-study (R10–15K within C1) What happens if a brand deliberately injects content designed to manipulate AI citations? Tests model robustness to prompt injection at the source layer. Publishable as a standalone methods paper.

Tier D — commercial scale-up (variable)

D1. White-label benchmark licensing Sell the methodology and code as a service to international partners running country-specific versions (Nigeria-AEO-Bench, Kenya-AEO-Bench, etc). Recurring license revenue.

D2. Real-time monitoring dashboard Continuous (daily/weekly) measurement of a focused 100-prompt subset for paying subscribers. Drift alerts when their brand citation share moves >5%. Subscription pricing per brand or per agency seat.

D3. API access to the dataset Expose the aggregate dataset (per-industry, per-model, per-domain citation counts) as a paid API. Marketing tech integrations.

Recommended sequence

Most efficient path to maximum value:

1. **Now (during data collection):** lock the v1 launch plan — PR list, embargo dates, GitHub repo polish, brief sponsor preview to lead targets.
2. **Week 1 post-collection:** publish v1 report + press launch (Tier A3). Captures the news cycle.
3. **Weeks 2–4 post-collection:** ship per-brand audits (Tier A1) and industry playbooks (Tier A2). Conversion-driven, while the news cycle is active.
4. **Months 2–3:** announce quarterly re-runs (Tier B1) as a subscription product. SMB v1 (Tier B2) goes into design.
5. **Months 4–6:** SMB v1 launches. Builds the “Cited Brands publishes regularly” muscle.
6. **Months 6–12:** SA-AEO-Bench v2 (Tier C1) launches at the 12-month anniversary as the annual flagship. R100K spend, ~10× the dataset size, tighter confidence intervals, multi-city, multi-language. Press cycle #3.

Risks and what could change the picture

1. **Streaming + real estate + restaurants haven't sampled meaningfully yet.** Streaming particularly might pull the ecommerce-bottom finding in a different direction (Netflix vs Showmax citations).
2. **Multilingual results are unknown.** If H5 fails (multilingual citations don't predominantly point to English content), the report shifts emphasis but doesn't break.
3. **Reddit asymmetry confirmation needs Vertex resolution to land at scale.** Spot-check at 500 URLs confirmed the pattern; full-run resolution will tighten the number.
4. **The dataset is timestamped May 2026.** LLM citation patterns drift over time. The v1 report needs to be published soon to preserve "current" relevance, and quarterly re-runs are the answer to drift.

Closing assessment

The dataset already supports a publishable, novel, commercially valuable report. Pre-registration on OSF gives it scientific credibility no commercial competitor has matched. Cost is on budget. Failure rate is healthy. Methodology assumptions are holding.

Eight of the findings listed in Section 4 are independently sellable as standalone insights. The data the run is still producing tightens these findings and adds three industries we haven't yet measured.

Stakeholder recommendation: **continue the run to completion, then publish the v1 report within 4 weeks of completion to capture the news cycle, with the per-brand audit and industry playbook products lined up for sale immediately after launch.**

This document will be updated when the run completes (~35% remaining). The final report is targeted for publication 4 weeks post-collection, on citedbrands.co.za/research/sa-aeo-bench-v1.