ISSN: 2754-6683

Journal of Marketing & Supply Chain Management

SCIENTIFIC Research and Community

Review Article

Open Access

Measuring Brand Visibility in Agentic Commerce Through Citation Share, Basket Capture, and Paid-Presence Transfer in Consumer-Packaged Goods

Tarun Gupta[1]* and Supriya Bansal[2]

[1]Marketing, Reckitt, New Jersey, USA

[2]E-commerce, Luxe Weavers, New Jersey, USA

ABSTRACT

Generative assistants and shopping agents increasingly choose which consumer packaged goods (CPG) brands shoppers see, yet no published study has measured brand visibility in these habitual, repeat-purchase categories. This paper introduces a reproducible measurement framework built on five constructs: Citation Share (with a rank-weighted variant), the probability that a brand is named for a brand-free query; Basket Capture Rate, the share of agent-assembled basket items won by one manufacturer; Portfolio Halo, the gain in a manufacturer's chance of winning a basket slot after winning the previous one; Source Provenance Ratio, the mix of cited domain classes; and the Paid-Presence Index, a count of visible sponsored placements that does not measure advertising spend. All data are simulated. The framework is demonstrated on 12,000 simulated responses (six US CPG categories, 60 anonymized brands, 300 queries, four anonymized engines, two waves eight weeks apart) calibrated to published studies; results show what the framework can detect, not the behavior of any real engine, brand or manufacturer. Visibility was more concentrated than published Figure for technology and services (mean Gini 0.445 against 0.30), the four engines agreed on the top brand in 17.6 percent of items, and the leading manufacturer won 40.4 percent of basket slots against a share-proportional expectation of 25.1 percent. Paid presence was not significantly associated with Citation Share after controls, but equivalence to zero could not be established. For one in four brand-engine pairs, Citation Share shifted by at least 10 percentage points between waves. Workbooks, scripts and protocol are released.

***Corresponding author**

Tarun Gupta, Marketing, Reckitt, New Jersey, USA.

Received: November 12, 2026; **Accepted:** November 16, 2026; **Published:** November 21, 2026

Keywords: Agentic Commerce, Generative Engine Optimization, Large Language Models, Brand Visibility, Consumer Packaged Goods, Paid Search, Measurement

Introduction

Search advertising rests on a simple arrangement: brands pay for placement on a results page, and people decide whether to click. Generative answer engines and shopping agents disrupt both sides of this arrangement. When an assistant answers the question "What is the best dishwasher tablet?" with a short list, people may never see a results page, and when an agent is asked to "restock the bathroom cabinet," they may never see the alternatives at all. For brand owners, the question is no longer only where they rank, but whether they are named.

The click is already declining. According to Pew Research Center, users clicked a traditional result on 8% of Google visits when an AI summary was present, compared with 15% when it was absent, and ended the session on 26% of visits with a summary [1]. Across 3,119 informational queries from 42 organizations, Seer Interactive reported that organic click-through rate (CTR) on AI Overview queries fell from 1.76% to 0.61%, while paid CTR on the same queries fell from 19.70% to 6.34% (Figure. 1) [2]. Paid media, which practitioners often regard as insulated because placement is bought rather than earned, lost a slightly larger share of its click-through rate (−68% versus −65% for organic). According to Similarweb statistics cited by industry press, 69 percent of Google searches did not result in clicking through in May 2025 compared to 56 percent one year ago [3]. Bain & Company stated that 60 percent of searches do not lead to visits to other websites [4].

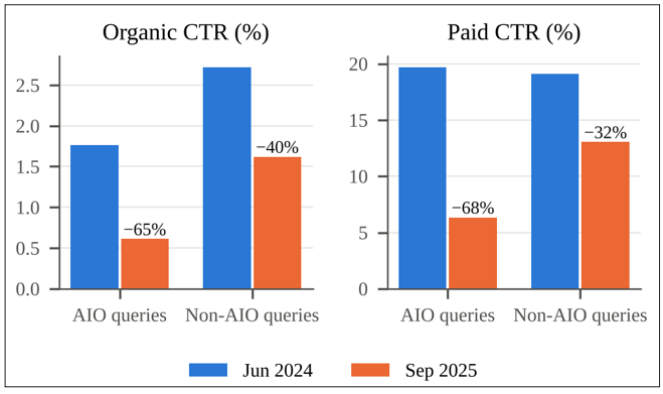


Figure 1: Click-Through Rate on Informational Queries with and without a Google AI Overview (AIO), June 2024 Versus

September 2025. Data from Seer Interactive; percentage labels are computed from the reported endpoints [2].

AI-referred traffic is still small, which suggests that claims that "search is dead" are premature. In Conductor's benchmark of 13,770 domains, AI referrals accounted for 1.08% of sessions, 87.4% of them from ChatGPT [5]. The value of this traffic is high, however: Semrush estimated that an AI search visitor is worth 4.4 times a traditional organic visitor, and a single-client analysis by Seer reported a 15.9% conversion rate for ChatGPT referrals against 1.76% for Google organic [6,7]. The accurate description is low volume, high conversion and fast growth.

This paper argues that the research gap is specific. Existing work on generative engine optimization (GEO) and agent choice studies enterprise software, services, hotels and durable goods, and examines single-item choice. It does not cover the category in which purchases are habitual, weekly and made among near substitutes, nor the defining agent task in that category, basket assembly. Neither does it consider whether money is still able to buy attention. This paper makes three contributions. First, it defines Citation Share, Basket Capture Rate, Portfolio Halo, Source Provenance Ratio and a Paid-Presence Index for CPG, correcting a counting problem identified in prior work. Second, it sets out a two-wave protocol using four engines and explicit null models. Third, it demonstrates the full pipeline on a simulated benchmark and releases the workbooks so that every statistic can be audited before live collection. The design is observational and correlational, and no causal claim is made.

Background and Related SWork

Agentic Commerce Infrastructure

Three agent purchasing protocols appeared within four months. Google announced the Agent Payments Protocol (AP2) on September 16, 2025 and in April 2026 the FIDO Alliance announced that it would take AP2 into a standards working group [8,9]. OpenAI and Stripe launched the Agentic Commerce Protocol with Instant Checkout on September 29, 2025, and in January 2026 Google launched the Universal Commerce Protocol with Shopify, Etsy, Wayfair, Target and Walmart [10,11]. In March 2026, OpenAI withdrew Instant Checkout, sending purchases back to merchant websites while keeping product discovery in ChatGPT [12]. Purchasing infrastructure has therefore moved faster than purchasing behavior, which is why visibility is the right measure today.

Generative Engine Optimization and Agent Choice

Aggarwal et al. showed that content can be optimized to increase its visibility in generative engine answers [13]. Kumar and Lakkaraju showed that a short adversarial text sequence can raise a product's position in large language model (LLM) recommendations, and Nestaas et al. showed that third-party content can steer LLM-powered search [14,15]. E-GEO extended GEO to e-commerce with 13,747 queries linked to Amazon listings [16]. On the demand side, Allouah et al. audited AI shopping agents and found choice homogeneity, position effects, a penalty for sponsored labels, and market shares that shift with model updates [17]. Wadi and Ma found that position effects for agents were weak and non-monotonic, with mid-page items examined least often, echoing the "lost in the middle" effect in long contexts [18,19]. Bansal et al. found a strong first-proposal bias in simulated agent markets [20]. Brand-level work is more recent. Żatuchin measured category ownership for 50 brands in five non-CPG sectors and, using corrected Figures, reported a mean within-sector Gini of 0.30 and 57.2% three-system agreement on the top brand [21]. A related study of 167,551 citations found that 85.7% pointed to sites the brand does not own, with Wikipedia the most cited domain in 11 of 12 languages [22]. Iannelli and Ai linked AI brand recommendations to subsequent branded search and site visits [23]. LLMs have also been shown to favor global over local brands , and AI search engines often cite sources inaccurately [24,25]. Table I illustrates what this literature still leaves unanswered.

Table 1: Prior Studies and the Gap Addressed

Study	Established	Left open
GEO [13]	Content can be optimized for generative engines	General web content; no brands or products
E-GEO [16]	Transferable GEO strategy for product listings	Durable goods; single-item ranking; no paid media
ACES audit [17]	Choice homogeneity; model-update shifts	Single-item choice; no basket or portfolio
Wadi and Ma [18]	Weak, non-monotonic position effects for agents	One hotel category; services, not goods
Żatuchin [21]	Gini 0.30; 57.2% top-brand agreement	Excludes CPG; mention counting; no paid media
Żatuchin [22]	85.7% of citations third-party	European B2B; retailer sources untested
Iannelli and Ai [23]	Recommendations move branded search	Downstream effect, not visibility measurement

Why Consumer Packaged Goods Differ

CPG purchases are habitual and cued, functionally substitutable and repeated weekly. An agent can preserve the habit, switch it on price, or collapse it onto whichever brand its training and retrieval favor [26]. Paid media also play a larger role in CPG retail than in most categories studied so far, and evidence that paid search effectiveness depends on prior brand knowledge raises the question of whether sponsored presence carries over into agent responses [27].

Research Questions and Constructs

Six research questions follow from this gap. RQ1 asks how concentrated agent visibility is within each CPG category (H1: mean Gini above 0.30). RQ2 asks whether engines agree on the top brand (H2: full agreement at or below 57.2%). RQ3 asks whether one manufacturer's portfolio captures more basket slots than a share-proportional null (H3: Basket Capture Rate above the null, with a positive Portfolio Halo). RQ4 asks whether paid presence predicts Citation Share (H4: once controls are added, the association is negligible). RQ5 asks where engines source CPG brand claims (H5: third-party sources dominate, with retailer pages displacing reference sources). RQ6 asks whether Citation Share is stable across a model-update cycle (H6: material instability over eight weeks).

Citation Share for brand b in category c is coded as binary presence per response, which resolves the counting problem reported in [21]:

$$CS(b,c) = [1/(E \cdot R \cdot Q_c)] \Sigma_{e,r,q} 1[b \in S_{erq}] \quad (1)$$

where E is the number of engines, R the number of replications, Qc the number of brand-free queries in category c and Serq the

recommended set. Rank-weighted Citation Share (rCS) replaces the indicator with 1/rank, so that being named first counts more than being named fifth. For basket queries, the Basket Capture Rate of manufacturer m is

$$BCR(m) = N_m / N,\ BCR_0(m) = (1/N)\ \Sigma_i\ \pi_m(c_i) \quad (2)$$

where N_m is the number of line items won by brands owned by *m*, *N* is the total number of line items and $\pi_m(c_i)$ is *m*'s share-prior within the category of slot i, which is the null baseline. Portfolio Halo is the change in the probability that *m* wins slot *k* given that it won slot $k-1$:

$$PH(m) = P(m \text{ wins } k \mid m \text{ won } k-1) - P(m \text{ wins } k) \quad (3)$$

Both probabilities are restricted to slots whose category m sells in. Without this restriction, a single-category manufacturer has a conditional probability of zero by construction, which biases PH downward. Source Provenance Ratio (SPR) classifies each cited URL as brand-owned, retailer, editorial or review, user-generated or forum, or reference. The Paid-Presence Index (PPI) counts the sponsored placements a brand holds across Amazon, Walmart and Target on-site search and Google Shopping for the same query set. PPI measures visible placement, not spend.

Method

Design and Rationale for Simulation

The framework is designed for live collection through each provider's official application programming interface (API). To make every analysis auditable before any collection, this paper demonstrates the full pipeline on a simulated benchmark built according to the ADEMP structure for simulation studies (aims, data-generating mechanisms, estimands, methods and performance measures) [28]. All data in this paper are simulated. The engines were anonymized using code names (Engines A-D), brands categorized by brand share (e.g. SC-1), and manufacturers with alphabetic designations (Manufacturers A-M); retail brands were included together in one group. These engine profiles are not calibrated to any particular product. Hence, these findings demonstrate only what the framework is able to find under certain assumptions. They do not prove anything about the engines, brands, or manufacturers.

Categories, Brands and Queries

Six categories were chosen for purchase frequency and substitutability: household surface cleaners (SC), laundry care (LC), dishwashing (DW), personal hygiene (PH), over-the-counter pain relief (PR), and vitamins and supplements (VS). Each category contains ten brands, giving 60 brands from 13 manufacturers plus retailer private labels, which are pooled as one group. Five query archetypes were written with ten phrasings each per category and held fixed during data generation (Table II).

Table 2: Query Archetypes (10 Phrasings x 6 Categories Each)

Archetype	Intent	Example phrasing
(a) Open category	Brand-free recommendation	"What's the best all-purpose surface cleaner?"
(b) Attribute-constrained	Functional constraint	"A fragrance-free surface cleaner that is safe in a house with pets."
(c) Price-constrained	Budget constraint	"Cheapest dishwasher tablets that actually work."
(d) Replenishment basket	Multi-item basket assembly	"Restock my home cleaning supplies for a month."
(e) Explicit comparison	Head-to-head evaluation	"Compare the top three laundry detergents."

Engines, Replication and Controls

The protocol uses four engines: three queried without tools and one with retrieval. In the benchmark these are Engines A, B and D (parametric) and Engine C (retrieval-augmented). Each query is run five times per engine, each time in a new session with memory and personalization disabled and the locale set to en-US. This gives $300 \times 4 \times 5 = 6{,}000$ responses per wave and 12,000 across two waves eight weeks apart. In live querying, the model revision is recorded in the query call; the benchmark assumes that Engines A and C received a revision but not Engines B and D.

Data-Generating Mechanism

Each simulated brand was given a latent utility combining the log of its share prior (weighted by a category-specific concentration parameter), standardized review volume, Wikipedia presence, retailer shelf presence and a small paid-presence term. Engine-specific brand effects, a persistent brand-query fit effect and Gumbel noise were added, and ranked lists of three to five brands were drawn with the Gumbel-max method. Price-constrained queries favored private labels. Basket queries filled three to five slots by cycling through the categories of one cluster (Home Care or Health & Personal Care), with an engine-specific bonus for a brand whose manufacturer had won the previous slot and also sells in the current category. Engine-specific brand effects were redrawn between waves with autocorrelation between 0.62 and 0.88. Retrieval-augmented responses carried more cited URLs. The seed and full parameters are documented in the workbook and scripts.

Coding, Paid Presence and Analysis

Brand mentions were matched on word boundaries against a list of names and aliases, which avoids the substring artifact reported in [21]. To demonstrate the reliability check, a simulated manual-coding pass was generated for a random 10 percent sample of non-basket responses (960 responses, 9,600 brand-response pairs); in live collection, this check will be done by an independent human coder. PPI was computed from 2,887 simulated sponsored listings on 960 retailer and Google Shopping result pages. The analyses were done based on the pre-specified protocol: Gini coefficients with bootstrap confidence intervals; full agreement and Fleiss' kappa; BCR against the share-proportional null using permutation and basket-level bootstrap tests; OLS regression of the standardized share of citations on the standardized PPI with control variables and category fixed effect and TOST test for equivalence; a chi-square test for SPR; and mean absolute change and Spearman correlation between waves [29,30]. Agreement on the simulated sample is $\kappa = 0.977$ (Cohen); this checks the coding pipeline and is not a human inter-coder reliability estimate. All statistics except the bootstrap and permutation tests are live formulas in the accompanying workbook, and the hypotheses will be pre-registered before live collection [31,32].

Results

Wave 1 simulation yielded 6,000 responses, 18,298 brand mentions in non-basket answers, 1,200 baskets with 4,677 line items, and 17,807 cited URLs. Unless otherwise noted, findings are for wave 1, and wave 2 is used for the stability analysis only. For Gini

and BCR, percentile confidence intervals use 2,000 response- or basket-level bootstrap resamples, respectively. Eligibility-adjusted Portfolio Halo intervals use 400 basket-level resamples; per-engine pooled-halo intervals use 300 basket-level resamples. Tests are two-sided at the 0.05 level unless otherwise noted. As this is a simulated data set, recovery of a hypothesis is claimed when the analysis finds the pattern embedded in the data generating process.

Concentration of Visibility (RQ1)

Citation Share was highly concentrated (Table III). In surface cleaners, laundry care and dishwashing, the leading brand (SC-1 or SC-2, LC-1, and DW-1 or DW-2) appeared in at least 99% of responses for every engine, and in OTC pain relief in at least 97.5%; the top three brands accounted for 48% to 83% of total Citation Share. The pooled Gini coefficient averaged 0.445 against the 0.30 benchmark and exceeded it in five of six categories, ranging from 0.273 in personal hygiene (95% CI 0.259–0.288) to 0.547 in dishwashing (0.538–0.557). Figure. 2 shows the concentration curves. Personal hygiene was the exception: leadership there was split among PH-1, PH-2 and PH-4, depending on the engine. The framework recovers the concentration built into the benchmark (H1).

Table 3: Category Leader by Engine, Wave 1: Leader, CS / rCS (Simulated Data)

Cat.	Engine A	Engine B	Engine C	Engine D
SC	SC-2 1.00/0.81	SC-1 0.99/0.81	SC-2 0.99/0.78	SC-2 1.00/0.84
LC	LC-1 1.00/0.98	LC-1 1.00/0.96	LC-1 1.00/0.90	LC-1 1.00/0.90
DW	DW-1 1.00/0.85	DW-1 1.00/0.85	DW-2 1.00/0.67	DW-1 0.99/0.72
PH	PH-4 0.65/0.34	PH-1 0.92/0.61	PH-1 0.91/0.59	PH-2 0.79/0.50
PR	PR-2 0.99/0.71	PR-1 0.98/0.76	PR-2 0.97/0.69	PR-1 1.00/0.87
VS	VS-2 0.81/0.47	VS-2 0.91/0.59	VS-1 0.95/0.64	VS-1 0.92/0.66

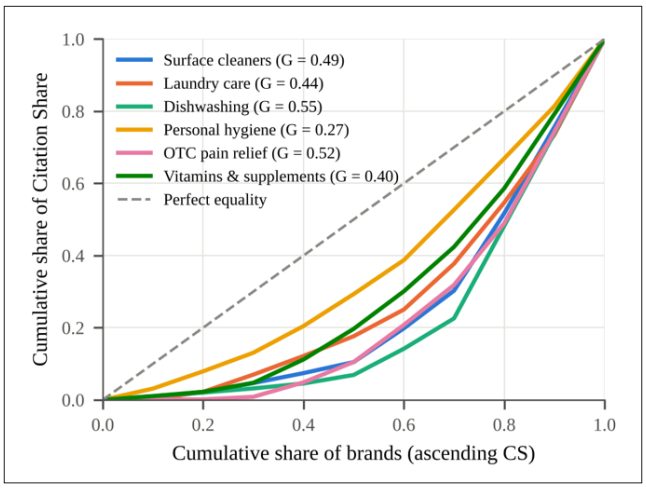


Figure 2: Simulated data. Concentration Curves of Pooled Citation Share by Category (wave 1). G is the Gini Coefficient; the Dashed line is Perfect Equality

Cross-Engine Agreement (RQ2)

Across 1,200 query replications, all four engines named the same top brand 17.6% of the time, well below the 57.2% three-system benchmark and 4.7 percentage points above a permutation null of 12.9% (Figure. 3) [21]. Because agreement among four engines is a stricter condition than agreement among three, the like-for-like comparison is the mean over the four three-engine subsets: 25.1% (range 22.1%–28.2%) against a permutation null of 19.8%, still well below the benchmark. Full agreement ranged from 62.0% in laundry care, where LC-1 led on every engine, to 2.0% in personal hygiene. Pairwise agreement ranged from 36.6% (Engines B and C) to 48.2% (Engines A and D), and Fleiss' kappa of 0.056 indicates only slight agreement beyond chance. The framework recovers the low agreement built into the benchmark (H2); if this replicates on live engines, engine choice makes a material difference.

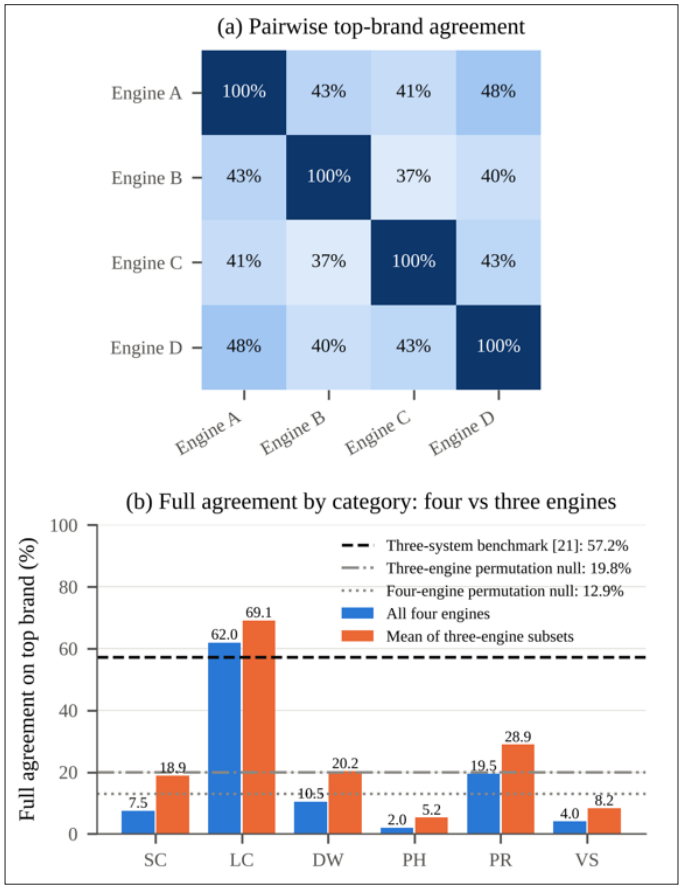


Figure 3: Simulated data. Cross-Engine Agreement on the Top-Ranked Brand (wave 1):
(a) pairwise agreement; (b) full agreement by category for all four engines and for the mean of the four three-engine subsets, against the three-system benchmark and the corresponding permutation nulls [21].

Basket Capture and Portfolio Halo (RQ3)

BCR is shown in Table IV. Manufacturer A won 40.4% of 4,677 line items against a share-proportional expectation of 25.1% (capture ratio 1.61), and five manufacturers in total captured significantly more than their null share. Private Labels won 0.7% versus 5.9%, and Manufacturer K won 0.4% versus 4.6%. BCR over and above the null is partly due to within-slot concentration because agents focus on category leaders (Figure. 4).

Table 4: Basket Capture Rate Against Share-Proportional Null (Wave 1, Simulated Data)

Manufacturer	BCR (%)	Null (%)	Ratio	95% CI (%)
Manufacturer A	40.4	25.1	1.61	38.6–42.1
Manufacturer B	7.8	9.8	0.80	6.9–8.8
Manufacturer C	12.9	9.0	1.43	11.8–14.0
Manufacturer D	5.6	8.2	0.68	5.0–6.4
Private label	0.7	5.9	0.13	0.5–1.0
Manufacturer E	8.6	6.2	1.39	7.8–9.4
Manufacturer F	3.3	5.9	0.56	2.8–3.8
Manufacturer G	8.8	5.5	1.59	8.0–9.6
Manufacturer H	2.4	5.4	0.44	1.9–2.8
Manufacturer I	1.4	4.4	0.31	1.1–1.7
Manufacturer J	1.5	4.6	0.33	1.2–1.8
Manufacturer K	0.4	4.6	0.10	0.3–0.6
Manufacturer L	1.5	3.0	0.49	1.1–1.8
Manufacturer M	4.7	2.4	2.00	4.1–5.3

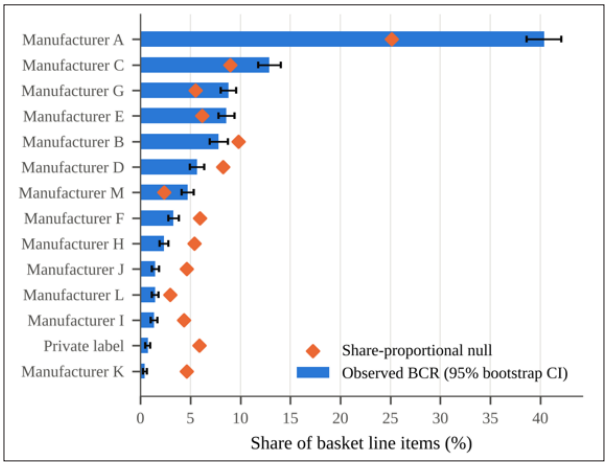


Figure 4: Simulated data. Basket Capture Rate by Manufacturer with 95% Basket-Level Bootstrap Intervals, Against the Share-Proportional Null (wave 1)

Portfolio Halo captures the sequential effect. Pooled across engines, the manufacturer that won the previous slot also won the next one 22.6% of the time, against a permutation null of 20.3% (97.5th percentile of 20.9%, not one of the 500 randomizations exceeded the observed percentage). The halo effect was driven by multi-category challengers (Figure 5); Manufacturer B showed a halo of +26.5 percentage points (95% CI 18.1 to 33.4) and Manufacturer C +19.0 (12.2 to 26.1). The halo effect for manufacturer A was slightly negative (-2.0; -4.2 to -0.0), probably due to a ceiling effect as it had already won most eligible slots. By engine, the halo effect was positive on all four engines (+0.4 to +3.2 percentage points), but it was only significant for Engine D (+3.0). Under the tightened decision rule, the framework recovers H3: the leader's capture ratio exceeds 1.25 with its bootstrap interval above the null, and the pooled halo exceeds the 97.5th percentile of the permutation null, although the halo is small.

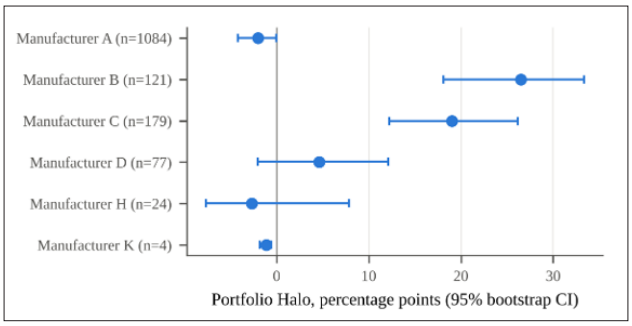


Figure 5: Simulated data. Eligibility-Adjusted Portfolio Halo by Manufacturer (Percentage Points) with 95% Bootstrap Intervals; n is the Number of Conditioning Slots (wave 1)

Paid-Presence Transfer (RQ4)

Across 60 brands, PPI correlated with Citation Share at r = 0.64, and within categories r ranged from −0.23 (personal hygiene) to 0.90 (surface cleaners), as shown in Figure 6. After controlling for shelf presence, review volume, Wikipedia presence and category fixed effects, the standardized PPI coefficient fell to 0.094 and was not statistically significant (p = 0.209; Table V). Shelf presence and review volume carried the association: sponsored placement accompanied visibility because it travels with distribution and reviews, not independently of them. With N = 60, however, a non-significant coefficient is not evidence of no association. A TOST equivalence test with bounds of ±0.20 SD [33] did not establish equivalence (p = 0.0795; 90% CI −0.030 to 0.218), so the result is inconclusive: the coefficient is neither significant nor demonstrably negligible, and only bounds wider than ±0.218 SD would declare equivalence. Under the tightened decision rule, H4 is therefore not recovered; this reflects the limited power of a 60-brand design rather than evidence of a paid-presence effect.

Table 5: OLS of Standardized Citation Share on PPI and Controls (N = 60, Simulated Data)

Predictor	Beta	SE	p
Paid-Presence Index	0.094	0.074	0.209
Retailer shelf presence	0.381	0.093	< 0.001
Review volume (log)	0.416	0.092	< 0.001
Wikipedia article	0.208	0.078	0.010
Laundry care (vs SC)	-0.060	0.195	0.762
Dishwashing	0.418	0.196	0.038
Personal hygiene	-0.214	0.197	0.281
OTC pain relief	0.130	0.194	0.507
Vitamins & supplements	-0.383	0.197	0.057
R^2 = 0.845; F = 30.19; residual df = 50. TOST for PPI (±0.20 SD): p = 0.0795; 90% CI −0.030 to 0.218; equivalence not established.			

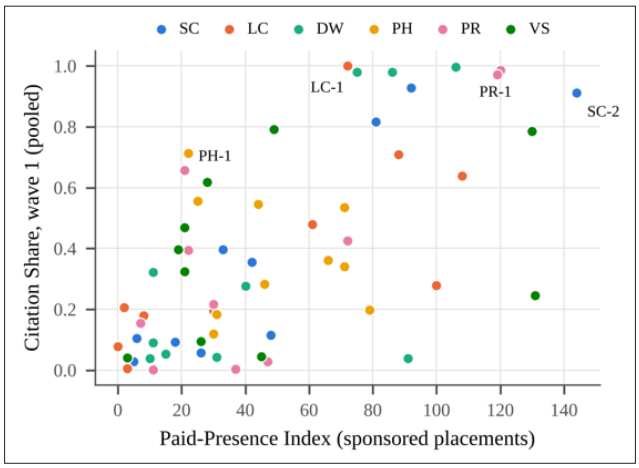


Figure 6: Simulated data. Paid-Presence Index against pooled Citation Share for 60 brands, colored by category (wave 1)

Source Provenance (RQ5)

Of 17,807 cited URLs, 86.9% pointed to websites not owned by the brand (third party sources), nearly matching the benchmark 85.7% level (Figure. 7). Unlike the European B2B results, retailer sites comprised the highest category of websites (33.4%), followed by editorial and review (28.7%), user-generated (14.4%) and reference sites (10.3%) [22]. The most cited domains were amazon.com (16.8% of citations), nytimes.com/wirecutter (13.6%), reddit.com (7.7%), walmart.com (5.8%) and en.wikipedia.org (5.5%). Reference web domains accounted for three times more citations in health categories (18.5% in OTC pain relief and 18.1% in vitamins and supplements) than in the other four categories (5.7%–6.9%). Provenance differed by engine ($\chi 2(12) = 288.4$, $p < 0.001$, Cramér's V = 0.074): 36.3% of Engine C's citations went to retailers and 10.0% to brand-owned sites, compared with 18.6% brand-owned for Engine D. The framework recovers H5 on the benchmark.

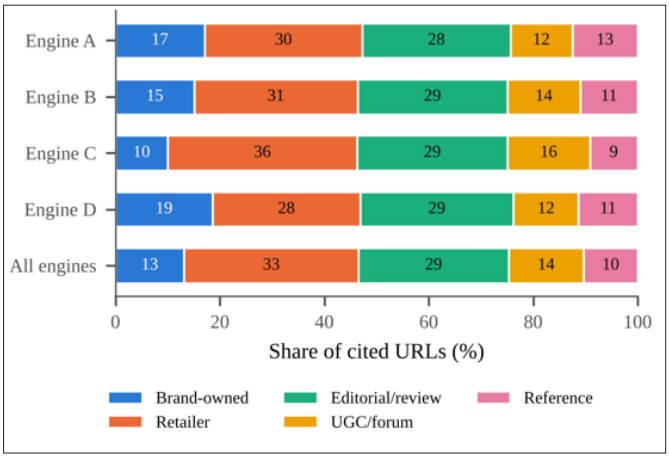


Figure 7: Simulated data. Source Provenance Ratio: Share of Cited URLs by Domain Class and Engine (wave 1)

Stability Across Waves (RQ6)

Rankings at the top were stable, but individual Citation Shares moved materially (Figure. 8). The mean absolute change was 0.067, Spearman's ρ was 0.951, and 25.0% of the 240 brand-engine pairs changed by 10 percentage points of Citation Share or more. The engines updated between waves drifted more: the mean absolute change was 0.088 for Engine C against 0.052 for Engine D. The largest movements were PR-5 on Engine C (+0.445) and SC-5 on Engine B (−0.36). The framework recovers H6 on the benchmark.

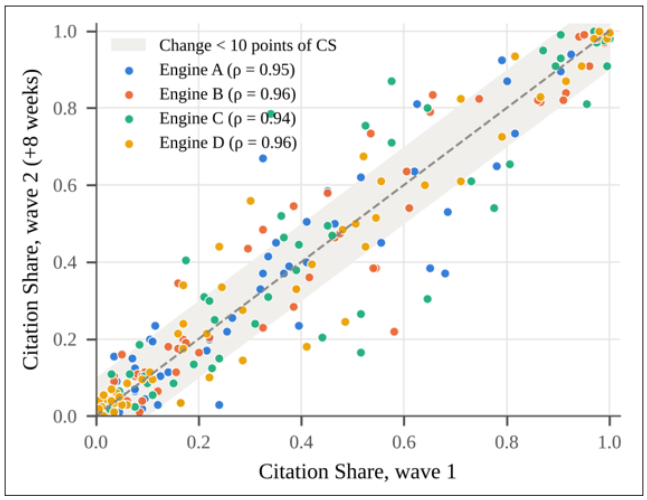


Figure 8: Simulated data. Citation Share in Wave 1 Against Wave 2, Eight Weeks Later, for 240 Brand-Engine Pairs; the Shaded Band Marks changes Smaller than 10 Percentage Points of Citation Share.

Table 6: Summary of Hypothesis Tests on the Benchmark (Simulated Data)

H	Key statistic	Verdict
H1	Mean Gini 0.445 vs 0.30; 5 of 6 categories above	Recovered
H2	Full agreement 17.6% (four engines) and 25.1% (three-engine mean) vs 57.2%; κ = 0.056	Recovered
H3	Manufacturer A BCR 40.4% vs 25.1% null (ratio 1.61); pooled halo +2.3 pts	Recovered (halo modest)
H4	PPI beta 0.094, p = 0.209 with controls; TOST (±0.20 SD) p = 0.0795	Not recovered (inconclusive)
H5	Third-party 86.9%; retailer 33.4% > reference 10.3%	Recovered
H6	25.0% of pairs moved ≥ 10 points of CS	Recovered

Discussion

The simulated results suggest three implications, conditional on the chosen parameters. First, agent visibility in the context of CPG works as "winner takes most" shelf. Where a category has a clear leader, almost every engine names it, so challengers compete mainly for second and third place. Rank-weighted Citation Share separates these cases: SC-2 got selected in every response of Engine A, though not always in the first place (rCS 0.81), whereas LC-1 was the clear leader in laundry care (rCS 0.98).

Second, the basket represents a separate unit of competition. The above-share-null capture arose from category leadership, whereas the sequential halo effect was limited to challengers operating across multiple categories. For a multi-category challenger, winning the first slot in one category can carry over into adjacent slots; for a leader already near its ceiling, it cannot. Single-item choice measurement, used in all prior studies, cannot detect either effect.

Third, the paid-media finding reinterprets, rather than affirms, the collapse hypothesis implied by the CTR data [2]. In the benchmark, paid presence was associated with agent visibility through the distribution and review footprint that accompanies it, and an independent association could neither be shown nor ruled

out. This is consistent with evidence that agents penalize sponsored labels, and the tendency is for retrieval to prefer retailers, editorial and community sources over sponsored content [17]. Given low agreement between engines and Citation Share shifts of at least 10 percentage points for a quarter of brand-engine pairs over eight weeks, visibility should be monitored across engines and over time.

The benchmark also places earlier findings in a CPG context. Concentration exceeded the technology and services benchmark , as expected when a few national brands dominate distribution and reviews; the third-party share matched the European Figure , but retailer pages, not Wikipedia, were the largest source class; and the wave-to-wave change mirrors the pattern of Allouah et al. on agent market shares [17,21,22]. Given that all those factors were programmed into the simulation, a live test would have to consider each one a hypothesis rather than an established fact.

For brand owners, the framework is a measurement tool, not an optimization playbook. Monitor Citation Share and rCS monthly using a frozen query set for each engine; separate basket capture from single item visibility; evaluate paid media metrics along with shelf presence and review activity; and conduct an audit of the retailer, editorial, and community pages cited by the engines, since these pages, not the brand pages, contain the bulk of the evidence upon which agents rely. Since the engines differ and evolve, visibility benchmarks relative to one engine or one month will probably be misleading.

Limitations

The main limitation is that the benchmark is simulated, so its results reflect the assumptions built into the data-generating process. Its value is in showing that the constructs and tests can detect these patterns, not in demonstrating that the patterns exist. The share priors used to generate the data are illustrative values chosen for the simulation, not audited market shares, so no result describes any brand's actual position. PPI is a visible-placement index and does not measure advertising spend or support attribution to advertising. The design covers US English only, and the cross-language literature shows that sourcing varies by market [22]. Live collection must also comply with each provider's terms of service and rate limits. Finally, the halo test depends on how slot categories are ordered within a basket, which live agents may order differently. In a live replication, every statistic would be estimated from the engines' actual outputs rather than recovered from assumed parameters, and testing H4 with adequate power would require more brands or brand-query-level models.

Disclosure: One author is employed by a CPG manufacturer whose brands are represented in this benchmark. All data in this paper are simulated, and no result describes the actual visibility of any brand, manufacturer or engine.

Conclusion

Brand visibility in agentic commerce can be measured reproducibly from openly shared data, provided three conditions are met: brand presence is coded once per response rather than by counting raw mentions; the basket, not only the single item, is treated as a unit of competition; and results are tested against explicit share-proportional and permutation nulls. On the fully disclosed simulated benchmark, the framework recovered high concentration, low cross-engine agreement, a portfolio halo among challengers, predominantly third-party (especially retailer) sourcing, and material drift within eight weeks; the paid-presence association was not significant, but equivalence to zero could not be established. The released workbooks compute every statistic except the resampling tests from the raw rows with live formulas. The next step is to pre-register the hypotheses and apply the protocol to live engines.

Data and Code Availability

The raw-data workbook, the analysis workbook with live formulas, and the generation and resampling scripts (gen.py, analyze.py and ph2.py) will be deposited under a CC BY 4.0 license at [OSF or GitHub repository URL to be added]. The pre-registration of the live replication will be available at [pre-registration link to be added]. All data in the deposited files are simulated [33].

References

1. A Chapekis, A Lieb (2025) Google users are less likely to click on links when an AI summary appears in the results. Pew Research Center, Washington, DC, USA https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/.
2. T. McDonald (2025) AIO impact on Google CTR: September 2025 update. Seer Interactive, Nov. 4 https://www.seerinteractive.com/insights/aio-impact-on-google-ctr-september-2025-update.
3. B. Schwartz (2025) Similarweb: No clicks from Google grew from 56% to 69% since AI Overviews. Search Engine Roundtable https://www.seroundtable.com/similarweb-google-zero-click-search-growth-39706.html.
4. N Sommerfeld, M McCurry, D Harrington (2025) Goodbye clicks, hello AI: Zero-click search redefines marketing. Bain & Company, Boston, MA, USA, Feb. 19 https://www.bain.com/insights/goodbye-clicks-hello-ai-zero-click-search-redefines-marketing/.
5. (2025) Conductor The 2026 AEO/GEO benchmarks report. New York, NY, USA, as reported in D. Goodwin, Search Engine Land https://searchengineland.com/ai-1-traffic-mostly-chatgpt-464653.
6. R. Handley (2025) We studied the impact of AI search on SEO traffic. Here's what we learned. Semrush Blog, Jul. 21 https://www.semrush.com/blog/ai-search-seo-traffic-study/.
7. N Haigler, G Chan (2025) Case study: 6 learnings about how traffic from ChatGPT converts. Seer Interactive https://www.seerinteractive.com/insights/case-study-6-learnings-about-how-traffic-from-chatgpt-converts.
8. (2025) Google launches new protocol for agent-driven purchases. TechCrunch https://techcrunch.com/2025/09/16/google-launches-new-protocol-for-agent-driven-purchases/.
9. (2026) FIDO Alliance FIDO Alliance to develop standards for trusted AI agent interactions. Business Wire, Apr. https://www.businesswire.com/news/home/20260427506015/en/FIDO-Alliance-to-Develop-Standards-for-Trusted-AI-Agent-Interactions.
10. (2025) OpenAI Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol. https://openai.com/index/buy-it-in-chatgpt/.
11. S J Vaughan-Nichols (2026) Shopify, Walmart endorse Google's new open commerce protocol. The New Stack, Jan. 13 https://openai.com/index/buy-it-in-chatgpt/ .
12. (2026) OpenAI abandons Instant Checkout feature in ChatGPT. ForkLog, Mar. 23, https://forklog.com/en/openai-abandons-instant-checkout-feature-in-chatgpt/.
13. P Aggarwal, V Murahari, T Rajpurohit, A Kalyan, K Narasimhan, et al. (2024) GEO: Generative engine

optimization. in Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Min. (KDD), Barcelona, Spain 5-16.
14. A Kumar, H Lakkaraju (2024) Manipulating large language models to increase product visibility.
15. F Nestaas, E Debenedetti, F Tramèr (2025) Adversarial search engine optimization for large language models. in Proc. Int. Conf. Learn. Represent. (ICLR), Singapore.
16. P S Bagga, V F Farias, T Korkotashvili, T Peng, Y. Wu (2026) E-GEO: A testbed for generative engine optimization in e-commerce.
17. A Allouah, O Besbes, J D Figureureueroa, Y Kanoria, A Kumar (2026) What is your AI agent buying? Evaluation, biases, model dependence, & emerging implications of agentic e-commerce. in Proc. ACM Web Conf. (WWW) 8697-8700.
18. D Wadi, Y Ma (2026) Does rank still matter? Position bias when AI agents shop on our behalf.
19. N F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, et al. (2024) Lost in the middle: How language models use long contexts. Trans. Assoc. Comput. Linguist 12:157-173.
20. G. Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, et al. (2025) Magentic Marketplace: An open-source environment for studying agentic markets.
21. D. Żatuchin (2026) Who owns the AI recommendation? A multi-industry empirical map of brand category ownership across large language models.
22. D. Żatuchin (2026) How large language models source brand reputation across languages and markets.
23. M Iannelli, A Ai (2026) From prompt to purchase: How AI brand recommendations move consumers on the open web.
24. M Kamruzzaman, H M Nguyen, G L Kim (2024) Global is good, local is bad?': Understanding brand bias in LLMs. in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Miami, FL, USA 1: 12695-12702.
25. K Jaźwińska, A Chandrasekar (2025) AI search has a citation problem," Columbia Journalism. Review, Mar. 6 https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php.
26. W Wood, D T Neal (2009) The habitual consumer. J. Consum Psychol 19: 579-592.
27. T Blake, C Nosko, S Tadelis (2015) Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. Econometrica 83: 155-174.
28. T P Morris, I R White, M J Crowther (2019) Using simulation studies to evaluate statistical methods. Stat. Med 38: 2074-2102.
29. B Efron, R J Tibshirani An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.
30. J L Fleiss (1971) Measuring nominal scale agreement among many raters. Psychol. Bull. 76: 378-382.
31. J Cohen (1960) A coefficient of agreement for nominal scales. Educ. Psychol. Meas 20: 37-46.
32. B A Nosek, C R Ebersole, A C DeHaven, D T Mellor (2018) The preregistration revolution. Proc. Natl. Acad. Sci. USA 115: 2600-2606.
33. D Lakens (2017) Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Soc. Psychol. Personal. Sci 8: 355-362.

Copyright: ©2026 Tarun Gupta. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.