ORBIE / RESEARCH

EVIDENCE REVIEW · OCTOBER 2026

The State of AI Search Visibility 2026

How AI systems retrieve, cite and recommend brands, what that does to demand, and how to measure it without fooling yourself. For marketers, SEO practitioners and agencies.

Orbie AI search team October 2026 11 studies reviewed Download PDF

A brand can be described accurately by an AI assistant and still never be recommended. A page can be retrieved and never cited. A recommendation can create demand that never shows up as an AI referral. These are different outcomes, and the research published in 2026 shows why treating them as one “ranking” gives a misleading picture of AI visibility.

This review brings together the strongest evidence available at the end of October 2026: two large academic studies of how ChatGPT retrieves and cites pages, the first critical survey of the field, controlled benchmarks of “GEO” content edits, a repeated-prompt experiment, a branded-citation audit, a behavioural panel with placebo controls, and two consumer surveys.

WHY IT MATTERS NOW

McKinsey’s June 2026 survey of 11 markets found that 53% of adults had used AI for at least one shopping task in the past three months (October 2026 analysis of about 21,000 consumers). That is self-reported use, not a share of purchases made inside AI tools. [11]

32% vs 2.8%

Share of within-site citation variation explained by what ChatGPT’s own search retrieved, against all tested page properties [1]

<1 in 100

Chance that ChatGPT or Google’s AI gives the same list of brands twice for the same prompt [3]

+4.3 pp

Rise in same-name Google searches after an AI assistant recommended a brand, over matched placebos (interval 3.1–5.5) [9]

KEY FINDINGS

What the evidence says, and what to do about it

1. Retrieval decides citation, not page formatting.

Within a website, what the engine's rewritten searches retrieve explains far more of citation than FAQ blocks, schema or page speed. [1] [2] **Do:** map the searches engines run for your questions, then make the page that should answer them match those searches.

2. Engines also name brands from memory.

ChatGPT writes brands it already knows into its searches and names some brands with no supporting page. [1] **Do:** treat how widely and consistently a brand is described across the web as part of visibility, not only its own pages.

3. Presence can be measured; position can't.

Lists and their order change almost every run, but the brands that appear often keep appearing. [3] **Do:** report "named in X of N answers" per engine, never a single rank or screenshot.

4. Generic "GEO tricks" transfer poorly.

Benchmarks show many content edits do nothing or harm when others adopt them; relevance and position in context are the reproducible levers. [4] [6] **Do:** test changes against a baseline instead of applying a checklist.

5. Demand shows up as branded search, not AI referrals.

After a recommendation, people search for the brand and visit its site; the path is mostly through search. [9] **Do:** track branded search and "how did you hear about us" next to AI visibility.

In this review

1. Four different visibility outcomes
2. Retrieval comes before formatting
3. Which sources AI relies on
4. Measuring recommendations
5. From recommendation to demand
6. A 90-day protocol for brands and agencies
7. Methods, limits and sources

“AI visibility” is four different outcomes

In classic search, analysts can see a ranked results page for one query at one moment. Generative engines add hidden steps: the assistant rewrites the question into its own searches, retrieves a set of candidate pages, cites some of them, and writes a recommendation that may name brands it never cites. Jiang and colleagues traced this process in four production ChatGPT Web settings (GPT-5.2 to 5.5, Instant and Thinking). Even combined, public Google results and the OpenAI Responses API recovered only about a quarter of the exact URLs ChatGPT cited. [2]

01 / FIND	02 / SUPPORT	03 / SELECT	04 / ACT
Retrieval	**Citation**	**Recommendation**	**Demand**
Does the page enter the engine’s candidate set for the searches it actually runs?	Does the engine visibly cite the page or domain?	Is the brand named as an option in the answer?	Does exposure lead to searches, visits or enquiries?

The four stages are a way to organise measurement, not a proven funnel: a brand can be recommended from model memory with no citation, and a citation does not guarantee a recommendation. External tools see final answers and citations; retrieval can be observed only where an engine exposes its searches, and inferred elsewhere. [1] [2]

WHAT THIS MEANS FOR PRACTITIONERS

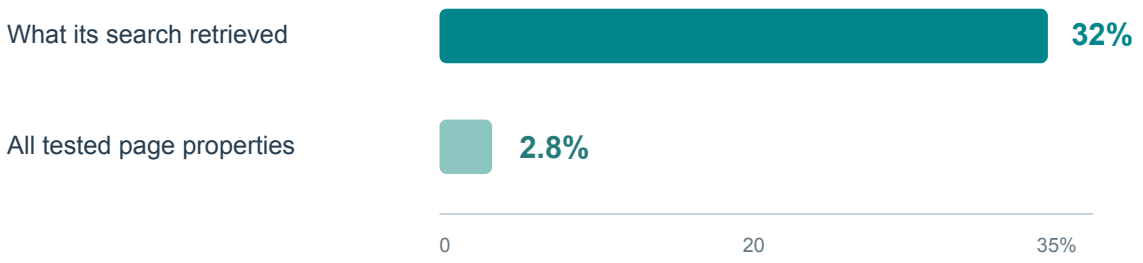
Report recommendation presence and citation share as separate numbers, with the engine, model setting, question set, number of runs and dates. A dashboard that merges them hides where a brand is actually losing.

The strongest new evidence points upstream of page formatting

In October 2026, Mahe Chen and Nick Koudas published a study of **158,099 answers and 468,037 citations** from two 90-day collections of commercial questions in contrasting markets. Its design is what makes it strong: by comparing pages *within the same website*, it separates the effect of a page feature from the general prominence of the site. [1]

Three results matter. First, FAQ blocks, structured data, summary boxes and page speed looked effective only when comparing different websites; within a site, page properties explained little. Pages whose titles and paragraphs matched the questions people ask were cited more. Second, when the authors replayed the questions through the engines' APIs, every page ChatGPT cited was one its own search had retrieved, and retrieval for the sub-queries the model rewrites on every answer explained **32%** of within-site variation, against **2.8%** for page properties. Those rewritten queries rarely matched Google's or Bing's results. Third, the model's own knowledge shapes the outcome: ChatGPT writes brands it knows into its searches, prefers websites it knows, names brands without any supporting page, and in controlled tests brand names alone reordered its rankings of products with identical facts. [1]

Figure 1 · What explains within-website citation variation (ChatGPT replay)



Source: Chen & Koudas (2026). Shares of explained variation in one ChatGPT replay analysis, not causal effect sizes or universal ranking weights. Chart drawn by Orbie from the reported figures. [1]

Jiang et al. add the production view. Instant settings mostly ran a single round of searches; GPT-5.4 Thinking often ran several. Pages retrieved more often across repeated runs were more likely to be cited. [2] The practical consequence: ranking in Google for the literal question is no guarantee of being retrieved, because the engine searches its own rewritten sub-questions.

What about the well-known “GEO” content edits?

The original GEO paper (Aggarwal et al., KDD 2024) reported visibility gains of up to about 40% from edits such as adding statistics, quotations and citations, in a benchmark where the source was already in the engine’s context. [5] It did not test whether those edits get a live page retrieved. C-SEO Bench (NeurIPS Datasets & Benchmarks 2025) tested such methods across two tasks, six domains and competing adopters: most were ineffective or harmful, improving a document’s position in the context worked better, and gains shrank as more competitors used the same methods. [6] Martinez’s critical survey of 45 studies reaches the same view: topical relevance and context position are the most reproducible levers, generic heuristics transfer poorly, and citation-oriented rewrites can even impair retrieval. [4]

> **Start from “make the right answer retrievable, attributable and true,” not from “write for the model.”**

WHAT THIS MEANS FOR PRACTITIONERS

Before changing templates, look at the searches the engine ran for the questions that matter and which pages they returned. The page that should be cited needs to answer those searches in plain terms. A schema or FAQ checklist on its own is not supported by the within-site evidence.

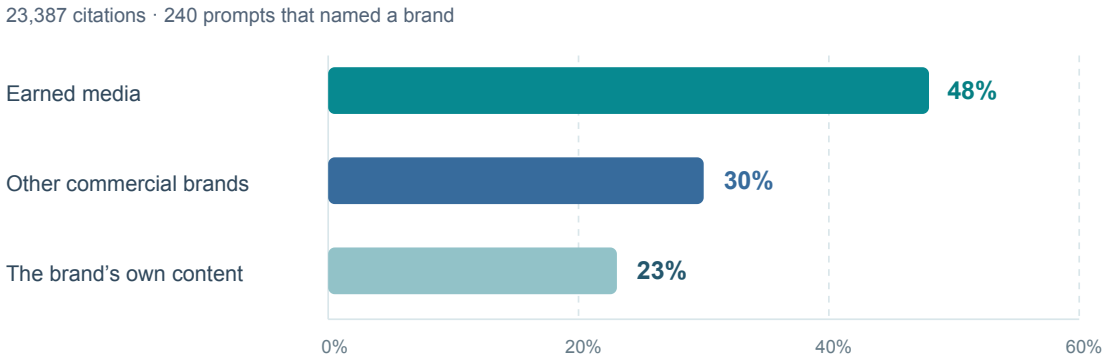
03 / SOURCES

Which sources AI relies on depends on the question

For an integration detail, the brand’s own documentation is the most direct evidence. For “which firm should I shortlist,” comparisons, directories and independent reviews carry more weight. A source-share percentage only means something with its denominator, engine and question type stated.

Omniscient Digital analysed **23,387 cited sources** from **240 prompts that named a brand**, across four industries (B2B SaaS, consumer products, professional services, finance) and six intent types. About 48% of citations were earned media, 30% other companies’ content and 23% the brand’s own content (rounded). Customer-review questions leaned on earned sources; functionality and integration questions used the brand’s own content more. [8]

Figure 2 · Source mix for prompts that name a brand (one study)



Source: Dombrowski / Omniscient Digital (January 2026). Rounded values total about 101%, so they are shown as separate bars. Branded-query evidence only; unbranded discovery questions will differ. [8]

Two cautions. These shares come from prompts that name the brand; unbranded "best X for Y" questions, where most new customers are won, can draw on a very different mix, and that mix varies by sector. And no study shows that appearing on a given third-party page produces a recommendation: pages differ in relevance, independence and which engines use them.

From backlinks to evidence coverage

A better question than "how many mentions do we have?" is "which buyer questions lack trustworthy evidence, and where does the evidence that engines use currently live?" For a pricing question, that is a current pricing page and corrected outdated listings. For a shortlist question, it is the directories, comparisons and reviews the engine actually cited. This question-to-evidence mapping is our working framework, built from the source patterns above; it is tested by re-measuring, not assumed.

Stale facts are a maintenance problem

Fang and colleagues attached artificial publication dates to otherwise identical passages and reranked them with seven language models. Fresher dates were consistently promoted, moving the top ten's mean publication year forward by up to 4.78 years and single passages by up to 95 ranks; larger models reduced the effect but none removed it. [7] The lesson is not to change dates. It is to keep the facts engines rely on current, on your own pages and on the external pages that describe you.

> **WHAT THIS MEANS FOR PRACTITIONERS**
>
> Audit sources per question type, not overall. For each important question, list the pages the engines cited, check what they say about the brand, and fix the wrong or missing facts on the ones that recur.

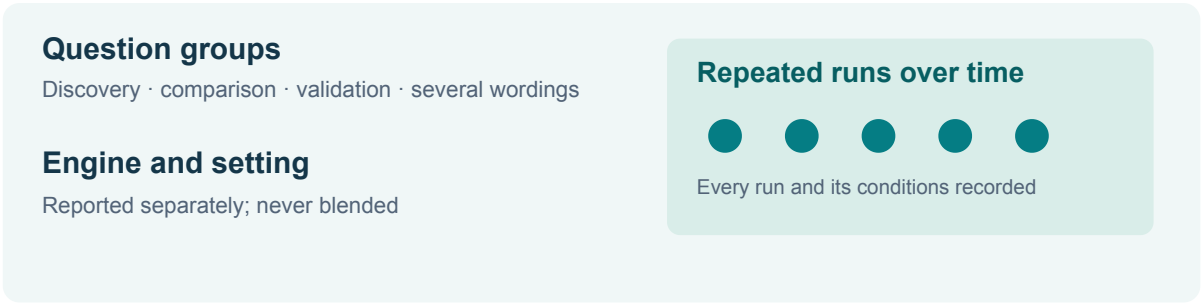
One AI answer is not a ranking, but presence can be measured

SparkToro's experiment (Fishkin and O'Donnell, January 2026) had 600 volunteers run 12 recommendation prompts 60–100 times each in ChatGPT, Claude and Google's AI, for 2,961 responses. ChatGPT and Google's AI had less than a 1-in-100 chance of giving the same list of brands twice, and about 1 in 1,000 of giving it in the same order. [3]

The same study found the opposite for presence. One cancer centre appeared in 69 of 71 ChatGPT answers while being the first mention in only 25. Across 994 responses to 142 different human-written headphone prompts, four brands appeared in 55–77% of answers. The authors conclude that visibility percentage across many prompts, run multiple times, is a reasonable metric, while ranking position is not. [3]

Put together, the evidence supports one way of measuring and rules out another. A position in one generated list is not a stable rank. The share of sampled answers that name or recommend a brand, per question group and engine, is a usable measure, provided the counts behind it are shown and each engine is reported separately.

Figure 3 · What a sound measurement design samples



Orbie's sampling design, informed by SparkToro [3] and Martinez [4]. Run counts depend on the question and the precision needed; no universal threshold has been established.

Counts matter because small samples mislead. A 40% recommendation rate from five runs is two answers; from 100 runs it is a far steadier estimate (though runs and wordings are not fully independent). Every headline percentage should come with its numerator, denominator, engine, setting, dates and question coverage.

Recommendation presence	answers that name or recommend the brand ÷ answers sampled
Owned-source citation rate	answers citing a brand-owned page ÷ answers that show citations
Accuracy rate	material claims about the brand that are correct ÷ claims checked

These are the definitions Orbie uses in its reports, not industry standards. A citation of the brand's domain is not the same as a recommendation of the brand.

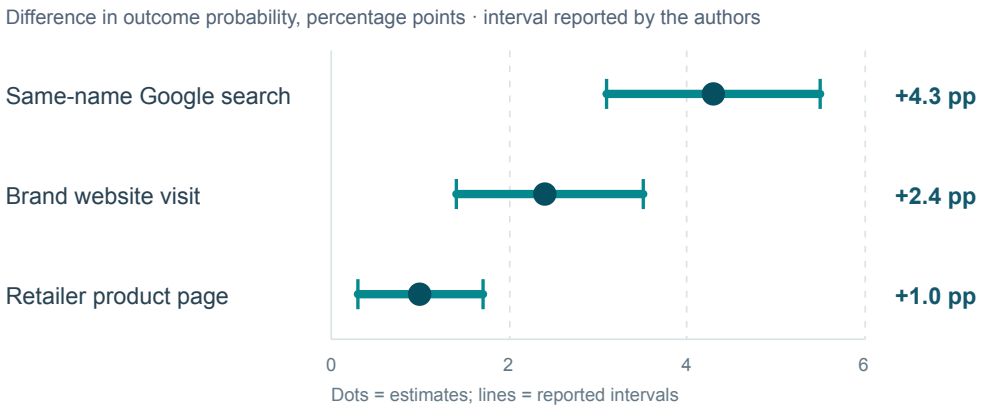
> **WHAT THIS MEANS FOR PRACTITIONERS**
>
> A single visibility score is useful only if it opens up into these parts: presence per question and engine, with the counts behind it. If a score can't be opened up, it hides the variation that matters. Agencies reporting to clients should show "named in X of N answers" per engine and keep screenshots as illustrations, not evidence.

05 / DEMAND

Recommendations lead to demand that referral reports miss

The best evidence so far on what a recommendation does comes from Iannelli and Ai (working paper, revised August 2026). They joined opt-in web browsing data with the same users' ChatGPT, Claude and Gemini conversations. Raw before/after comparisons overstate the effect, because many mentions are incidental references to brands the user already uses. Using matched backward placebos, a stance classifier, non-customer conditioning and a same-category control within each response, they estimate that a recommendation to a user with no recent engagement with the brand raises same-name Google searches by **4.3 percentage points**, visits to the brand's site by **2.4**, and brand-specific retailer-page visits by **1.0**. Incidental name-drops moved behaviour far less (+1.8 / +1.1 / +0.3). The path ran mostly through search. [9]

Figure 4 · Behaviour after an AI recommendation, over matched placebos



Source: Iannelli & Ai (2026), *From Prompt to Purchase*. Observational estimates, not randomised effects; purchases were not observed. Chart drawn by Orbie from the reported figures. [9]

Two consequences follow. AI referral traffic in analytics captures only part of the effect, because most people who act on a recommendation search for the brand rather than click a link. And no single branded-search increase can be credited to one placement or page change without a comparison. McKinsey's surveys give the context: in its five-country survey of 4,863 consumers, 28% of Gen Z and 16% of baby boomers said they used generative AI for shopping. [10] [11]

> **WHAT THIS MEANS FOR PRACTITIONERS**
>
> Measure AI visibility next to branded search trends, a "how did you hear about us" option and enquiry quality. Visibility is an intermediate measure; the business question is whether better visibility is followed by better demand.

06 / APPLICATION

A 90-day protocol for brands and agencies

This is the protocol Orbie follows, derived from the evidence above. The numbers in it are practical defaults, not validated thresholds. Agencies can run it per client and report it in the same format.

DAYS 1–30 / BASELINE

Define the questions, measure, and record everything

- Choose 20–40 questions real buyers ask: category discovery, alternatives, comparison and verification, each in at least two natural wordings, scoped to the market the business actually serves.
- Run each question several times per engine over several days. Keep raw answers, timestamps, the searches the engine ran, cited URLs, and the answers where the brand is absent.
- Record mention, recommendation, citation, cited domain and factual errors as separate fields.
- List the facts that matter (prices, services, locations, credentials) and find where external pages contradict them.

Deliverable: a baseline with counts per question group and engine, and a register of sources and wrong facts.

DAYS 31–60 / FIX

Close the evidence gaps that the baseline shows

- For each question where the brand is missing, identify the likely cause: the engine's searches don't reach the right page, the page doesn't answer them, independent evidence is thin, or the facts are unclear.
- Improve one page or page group at a time so it answers the searches the engine ran, in plain terms, with the entities named.
- Where recurring third-party sources are wrong or leave the brand out, ask for corrections or a fair inclusion; disclose any sponsorship and never manufacture reviews.
- Log each change with the question group it should affect and the result that would show it didn't work.

Deliverable: a change log with dates, URLs, expected effects and untouched comparison questions.

DAYS 61–90 / EVALUATE

Re-measure and separate signal from noise

- Re-run the same questions on the same engines; keep a group of untouched questions as a comparison. Compare presence and citation rates with their intervals, not rank changes.
- Check that changes hold across wordings, runs and a later date; note platform updates, launches and campaigns that could explain them.
- Set visibility changes against branded search, enquiries and what customers say about how they found the business.
- Report what didn't work as well as what did.

Deliverable: a before/after readout with counts, intervals and the next tests.

Example: a commercial law firm

A firm is not named when people ask for an employment solicitor for small businesses in its city. First check whether engines describe it correctly when asked about it by name: if not, the problem is basic recognition, not ranking. Then look at the searches the engines ran and the pages they cited: typically legal directories and rankings, review sites and rival firms' service pages. If the firm's own employment page doesn't state who it serves and where, fix that first; if the directories that keep being cited have an outdated or missing profile, fix those. Re-measure the same questions and report "named in X of N answers" per engine, before and after.

Example: an agency reporting to a client

An agency adds AI visibility to a monthly client report. Instead of a screenshot of one answer, it shows presence per question group and engine with the counts behind it, the sources cited most often for the client's questions, the facts that are wrong in those sources, and the changes made that month. Movements are only reported as gains when they exceed run-to-run variation and hold on the untouched comparison questions.

Measure this for a brand or a client

Orbie measures how ChatGPT, Gemini and Perplexity recommend and describe a business, with the method in this review: repeated runs, results per engine, the sources behind each answer, and actions checked by a person.

- See a sample report
- For agencies: white-label reports for your clients
- Request a free light report for your brand

07 / METHODS

How this review was built, and how far to trust it

Scope: 11 published studies (2024–2026) on retrieval, source selection, recommendation variability and downstream behaviour. Sources were found through arXiv searches, the reference list of Martinez's 2026 survey [4] and public literature reviews. Every figure quoted is taken from the original publication, which we read; none is taken from a secondary summary.

Selection: we preferred peer-reviewed papers, preprints and studies whose method is public, and left out studies whose headline figures cannot be checked against a published method. Where authors have a commercial interest in the result, the register says so. This is a purposeful, non-systematic review, not a meta-analysis: it does not pool effects.

Higher confidence: tested mechanisms

Controlled studies show how documents already in context are used or reranked under given settings. They may not carry over to live engines.

Higher confidence: descriptive patterns

Large citation datasets describe their engines, questions and dates well. Scale does not remove selection bias.

Directional: business effects

Panels show behaviour after exposure; opt-in samples and confounding limit causal claims.

Low confidence: universal "GEO hacks"

Claims of guaranteed ranking or fixed mention thresholds would need replicated interventions on live engines.

Evidence register

Source	Design / scope	Used for	Main limitation
Chen & Koudas, Oct 2026	158,099 answers; 468,037 citations; two 90-day commercial collections + API replay	Retrieval vs page properties; model knowledge	Preprint; two markets; API settings
Jiang et al., Oct 2026	Repeated production traces, four ChatGPT Web settings (GPT-5.2–5.5)	Retrieval-to-citation; proxy mismatch	Preprint; ChatGPT only
Fishkin & O'Donnell (SparkToro), Jan 2026	600 volunteers; 2,961 responses; 12 prompts; follow-up of 142 prompts / 994 responses	List and order variation; presence stability	Small prompt set; a co-author works for an AI-visibility vendor
Martinez, Jul 2026	Critical survey of 45 GEO studies	Evidence boundaries; reproducible levers	Depends on the reviewed literature
Aggarwal et al., KDD 2024	GEO-bench; controlled content edits	What content edits do in a fixed context	Benchmark, not live discovery; older models
Puerto et al., NeurIPS D&B 2025	C-SEO Bench; 2 tasks, 6 domains, competing adopters	Limits of content edits; context position	Benchmark, not a live field trial
Fang et al., 2025 (rev. Sep 2026)	Artificial dates; seven reranking models	Recency bias	Artificial dates; no traffic outcome
Dombrowski (Omniscient Digital), Jan 2026	23,387 cited sources; 240 branded prompts; 4 industries, 6 intents	Source mix by question type	Brand-named prompts only; agency study
Iannelli & Ai, Jun 2026 (rev. Aug 2026)	Opt-in clickstream joined to ChatGPT, Claude and Gemini conversations; placebo-controlled event study	Behaviour after a recommendation	Observational; no transactions; first author works for an AI-visibility vendor
McKinsey, Jun 2026	4,863 consumers; five countries	Context: AI use for shopping by age	Self-report
McKinsey, 28 Oct 2026	About 21,000 respondents; multiple markets	Context: AI use for shopping	Self-report; not comparable with [10]

References

1. Chen, M. & Koudas, N. (2026). *What Drives AI Citations? Content, Retrieval and Model Knowledge in Generative Engine Optimization (GEO).* Preprint. arxiv.org/abs/2610.15033
2. Jiang, Y., Zhang, S., Xing, A., Chang, X. & Xiao, X. (2026). *Understanding ChatGPT Citations: Evidence from Production Web Search Across Model Generations.* Preprint. arxiv.org/abs/2610.15382
3. Fishkin, R. & O'Donnell, P. (28 January 2026). *NEW Research: AIs are highly inconsistent when recommending brands*

or products; marketers should take care when tracking AI visibility. SparkToro. sparktoro.com

4. Martinez, O. (2026). *Optimizing Visibility in Generative Engines: A Critical Survey of Generative Engine Optimization (2023–2026).* Preprint. arxiv.org/abs/2607.14035
5. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K. & Deshpande, A. (2024). *GEO: Generative Engine Optimization.* KDD 2024. arxiv.org/abs/2311.09735
6. Puerto, H., Gubri, M., Green, T., Oh, S. J. & Yun, S. (2025). *C-SEO Bench: Does Conversational SEO Work?* NeurIPS Datasets & Benchmarks 2025. arxiv.org/abs/2506.11097
7. Fang, H., Tao, S., Chen, N., Chang, K.-X. & Sakai, T. (2025). *Do Large Language Models Favor Recent Content? A Study on Recency Bias in LLM-Based Reranking.* arxiv.org/abs/2509.11353
8. Dombrowski, C. (8 January 2026). *How LLMs Source Brand Information: An Analysis of 23,000+ AI Citations.* Omniscient Digital. beomniscient.com
9. Iannelli, M. & Ai, A. (2026). *From Prompt to Purchase: How AI Brand Recommendations Move Consumers on the Open Web.* Working paper, v2. arxiv.org/abs/2606.10907
10. McKinsey & Company (22 June 2026). *State of the Consumer 2026: When Tech Acceleration and Cost Pressures Collide.* mckinsey.com
11. McKinsey & Company (28 October 2026). *Who Wins When AI Does the Shopping? New Evidence on Agentic Commerce.* mckinsey.com

> **Publication note.** Orbie did not run these experiments or collect these panels. Preprint findings may change after peer review. Sources last checked on 3 November 2026.