Cited or Invisible? A Multilingual, Multi-Platform Audit of Brand Citation in Generative Search

Muhammad Arshad

Independent researcher · arshad@telictech.com

6 September 2026

Abstract

Generative search engines increasingly answer commercial questions directly, citing a handful of sources rather than returning a ranked list. Which businesses those systems cite is becoming consequential for market access, yet the question is mostly discussed through vendor marketing rather than measurement. We report two audits of brand citation across four generative answer surfaces (Google AI Overviews, Gemini with search grounding, Perplexity, and ChatGPT with web search) in six European languages. Prompts are authored natively in each language rather than translated, because translation measures the translator and not the market, and each is issued three times per platform. Study 1 covers five commercial verticals with a set of replica furniture retailers as focal brands (4,320 observations); Study 2 covers authenticity-seeking queries with eight design rights holders as focal brands (1,080 observations). Neither run had a failed call.

Four findings. First, the platform a question is asked on dominates the language it is asked in, and the effect replicates across two independent brand sets: citation rates run 16.2% to 91.2% across platforms in Study 1 and 6.3% to 63.3% in Study 2, with the same platform ordering in both, while the spread across six languages is 41.0–61.1% and 31.1–46.1% respectively with intervals that overlap almost throughout. Second, we show by resampling how readily prompt selection manufactures a language gap: an eight-prompt single-pass study reports a difference of at least 20 points 16.4% of the time when the corpus-wide difference is 8.3 points, and 12.5% of the time on a platform where the true difference is exactly zero. Our own pilot produced a 38-point gap by this mechanism. Third, repeated identical queries disagree with themselves at rates differing by more than an order of magnitude across platforms, and that ordering also replicates. Fourth, on queries explicitly seeking the authentic article, Google AI Overviews cite the actual rights holder in 6.3% of answers and produce an overview at all for 30% of queries; a replica retailer is the fourth most-cited domain, ahead of four of the eight rights holders.

We release the measurement harness, the prompt corpora, and the observation logs. Total data-collection cost for the reported studies was $57.65.

1. Introduction

Search is shifting from a list of links to a synthesised answer that names a few sources. For an informational query this is a convenience. For a commercial one — *which sauna should I buy*, *where do I buy a genuine Eames chair* — it is closer to an allocation decision: the businesses named are the businesses a user will consider, and the rest are not merely ranked lower but absent.

A commercial category has grown up around this, selling "AI visibility" monitoring to firms worried about that absence. Its central empirical claims — that visibility differs sharply by language, that adding a language multiplies citation rates several-fold — are repeated widely and, as far as we can establish, rest on small or undisclosed samples. That matters beyond marketing. If the claims are right, non-

English businesses face a structural disadvantage in how AI systems mediate markets. If they are artifacts, firms are being sold a remedy for a problem they may not have.

This paper measures the question directly, and measures it twice on disjoint brand sets so that the result does not depend on the idiosyncrasies of any one set of websites. Our contributions are:

1. **A cross-platform, cross-language citation audit, replicated.** Four answer surfaces, six languages, two independent focal brand sets, with replicates, released in full. Concurrent work (Kumar, 2026) measures brand visibility at larger scale but does not vary the language of the prompt, does not replicate, and releases neither data nor code.
2. **A demonstration that prompt selection can manufacture a language gap.** A pilot of ours produced a 38-point English–German gap that fell to four points when the sample was widened, with no other change. We quantify how large an apparent effect prompt choice alone can produce.
3. **An instability measure**, treated as a first-class quantity rather than noise to be averaged away, and shown to replicate in ordering across both studies.
4. **Evidence that authenticity-seeking demand is not routed to rights holders.** On queries explicitly asking where to buy the genuine article, the manufacturers who own the designs are cited rarely, and a replica retailer outranks most of them.
5. **Documented extraction pitfalls.** We report two citation-parsing failures, one of which silently produces a citation rate of exactly zero, and which we initially mistook for a finding.

2. Related work

Algorithm auditing. Our design follows the external-audit tradition established by Sandvig et al. (2014) and surveyed by Metaxa et al. (2021): probe a deployed system from the outside with controlled inputs, treating it as a black box whose behaviour is measured rather than explained.

Generative search and verifiability. Liu et al. (2023) evaluate whether generative search engines' citations support the claims attached to them, and find substantial unsupported attribution. We ask a complementary question — not whether a citation is faithful, but *whose* content gets cited at all.

Generative engine optimisation. Aggarwal et al. (2024) study how content can be modified to increase the chance of being cited. That literature takes the perspective of a site trying to be included; we take the perspective of an auditor asking how inclusion is distributed.

Concurrent large-scale measurement. Kumar (2026) reports a baseline over 100,000 prompt responses and more than 100 brands, finding that brand visibility is measurable, differs by platform, and varies with brand maturity. That study is an order of magnitude larger than ours in responses and we do not attempt to improve on its scale. We differ on axes scale does not address: it does not appear to vary the language of the prompt, reports single observations per prompt, and releases neither code nor data, so its numbers cannot be recomputed.

Auditing what gets cited. Allaham and Diakopoulos (2026) audit the same four engines, asking whether they cite AI-generated sources, and find roughly 16% of cited sources to be synthetic. Their question concerns the *nature* of the cited source; ours concerns *whose* source is cited.

3. Method

3.1 Two studies

Study 1 samples the cross-product of six markets and five commercial verticals: design-furniture (replica retail), home-wellness (saunas and infrared cabins), local-services (city-anchored SMEs), b2b-software (domestic incumbents against US brands) and outdoor-garden. Twelve prompts per vertical per market gives 360 prompts. Focal brands are four replica furniture retailers, reported pseudonymously (§3.6).

Study 2 was added specifically to check whether Study 1's findings depend on its focal brands. It uses a new corpus of fifteen prompts per market — 90 in total — written for a buyer seeking the *authentic* article rather than a replica, and takes as focal brands eight rights holders: Vitra, Herman Miller, Knoll, Cassina, Carl Hansen & Søn, Fritz Hansen, Louis Poulsen and Flos. These brands share no ownership, no SEO practice and no commercial relationship with Study 1's focal set.

3.2 Prompt authoring

Prompts are authored natively in each language and are *never* translated between markets. This is a methodological commitment, not a convenience. German buyers of replica furniture search *Nachbau*; Dutch buyers say *namaak*; neither is what a dictionary returns for "replica". A translated corpus measures the translation pipeline and reports the result as a property of the market. We regard translated multilingual prompt sets as a threat to validity in any study of this kind, including several whose figures we set out to check.

3.3 Platforms and access mode

We query Google AI Overviews (retrieved geo- and language-pinned via a SERP data provider), Gemini with Google Search grounding, Perplexity Sonar, and OpenAI's Responses API with the web-search tool.

A limitation we state rather than bury. API responses are not identical to what a person sees in the corresponding consumer product. Personalisation, model routing, interface-level re-ranking and product-specific retrieval all differ. Scraping the consumer interfaces would breach the providers' terms of service, so every study in this area faces the same choice; many do not disclose which side of it they are on. Our measurements describe the API surfaces. Where a platform exposes a country parameter we set it to the market's country.

3.4 Citation extraction

A response *cites* a domain when that domain appears in the structured citation list the provider returns. We deliberately do not count a brand named in prose but not linked; naming and citing are different acts, and conflating them is how a visibility score becomes unfalsifiable.

A trap worth documenting. Gemini returns grounding chunks whose `uri` field is an opaque redirect through a Google-owned host; on the API version we used there is no domain field at all, and the source host appears only in the chunk *title*. An extractor that reads `uri` — the obvious choice — resolves every citation to Google and reports a citation rate of exactly zero, in every language, for every brand. We made this error, recorded "Gemini cites almost nobody" as a finding, and only caught it on manual inspection of a response that visibly did contain the brand.

A second fault is easy to reproduce. Our domain normaliser initially truncated hostnames under two-part public suffixes, turning `shop.example.com.au` into `com.au`. This affected 139 citations across 3.0%

of Study 1 observations. It touched no focal-brand rate, every focal domain being two-label, but introduced 22 spurious rows into the concentration analysis; the released data excludes them and reports the count.

3.5 Replicates and statistical treatment

Every prompt is issued three times per platform. Generative answers are not deterministic, and with one observation per cell there is no way to separate a real difference from resampling variation. Confidence intervals come from a cluster bootstrap that resamples *prompts*, not observations; three replicates of one question are not three independent questions.

Focal-brand analyses are restricted to the vertical in which the focal brands trade. Pooling in verticals where a brand sells nothing adds a population whose true citation rate is zero and whose replicates are therefore always unanimous; in Study 1 this deflated measured instability by roughly a factor of five.

3.6 Pseudonymisation and disclosure

Study 1's four focal domains, and one domain appearing in Study 2's concentration results, are operated by the author and are reported as **R1–R4**. This is a redaction for legal reasons rather than a methodological choice, and we state it plainly because a silent omission would be worse: R1 in particular is one of the most-cited domains in Study 2 (§4.5), and suppressing that without saying so would misrepresent the result. No reported quantity depends on the identity of these domains; every one is a function of the citation outcome. Readers who wish to verify the analysis against named domains should note that the released Study 2 log is unredacted apart from these entries.

3.7 Cost accounting

Every API call is logged with its cost. Study 1 cost $46.14 and Study 2 $11.50, a reported total of $57.65. We also report the cost of a discarded pilot ($14.07), because a study that reports only its successful run misrepresents the cost of doing the work.

4. Results

4.1 Platform dominates language, in both studies

Table 1 gives the share of answers citing a focal domain in each study. The platform ordering is identical across two disjoint brand sets and two disjoint prompt corpora: Perplexity highest, then Gemini, then ChatGPT, then Google AI Overviews, which is lowest by a wide margin in both.

Platform	Study 1	95% CI	Study 2	95% CI
Perplexity	91.2%	[84.3–97.2]	63.3%	[53.3–73.0]
Gemini (grounded)	63.9%	[55.1–72.7]	54.4%	[45.6–63.3]
ChatGPT (web search)	36.1%	[25.5–46.8]	30.4%	[21.9–38.9]
Google AI Overviews	16.2%	[9.3–24.1]	6.3%	[3.0–10.0]

Table 1. Citation rate by platform. Study 1: replica retailers, design-furniture vertical, n = 216 per platform. Study 2: rights holders, authenticity queries, n = 270 per platform.

In Study 1 no pair of intervals overlaps. In Study 2 the Perplexity–Gemini pair overlaps and we do not claim separation there; every other pair separates. The range from most to least citing platform is a factor of 5.6 in Study 1 and 10.0 in Study 2.

Market	Study 1	95% CI	Study 2	95% CI
English	58.3%	[41.7–73.6]	46.1%	[32.8–58.9]
German	52.1%	[36.1–68.1]	42.8%	[27.2–58.3]
Spanish	61.1%	[54.2–68.1]	39.4%	[26.7–52.2]
Dutch	48.6%	[42.4–55.6]	37.8%	[25.0–50.6]
French	41.0%	[30.6–50.7]	34.4%	[23.3–46.1]
Italian	50.0%	[35.4–63.2]	31.1%	[19.4–42.8]

Table 2. Citation rate by market. Study 1 n = 144 per market; Study 2 n = 180 per market.

Across languages the intervals overlap almost throughout in both studies. In Study 1 only the Spanish–French pair separates at the margin, and by the power analysis in the companion paper a design with twelve prompts per market is underpowered to confirm a difference of that size; we therefore do not claim it. In Study 2 no language pair separates at all. The headline comparison is between a platform effect that is large, ordered, replicated and unambiguous, and a language effect that neither design can resolve. That is the opposite of the emphasis in the commercial literature.

4.2 Answer availability is not universal

Three of the four platforms produced an answer to every query in both studies. Google AI Overviews produced one for 38% of Study 1 queries and 30% of Study 2 queries. On the remainder the question of citation does not arise — a fact that any visibility metric averaging over "queries" rather than "answers" will silently absorb.

4.3 How a language gap can be manufactured

We resample Study 1 under the conditions in which language-gap claims are typically produced: one platform, one pass, a handful of prompts per language. For each subsample size k we draw k matched question slots, compute the English–German difference, and repeat 5,000 times.

Conditions	True gap	P(≥20pt)	Max draw
Google AIO, k = 4	8.3	33.5%	75.0
Google AIO, k = 8	8.3	16.4%	50.0
Google AIO, k = 25	8.3	2.7%	24.0
ChatGPT, k = 4	0.0	12.5%	25.0
ChatGPT, k = 12	0.0	0.0%	8.3

Table 3. Apparent English–German gap under single-pass conditions.

On ChatGPT, where the corpus-wide difference is exactly zero, a four-prompt study still returns a difference of at least 20 points in one direction or the other 12.5% of the time. This is not hypothetical: our own pilot, on eight prompts per language, measured a 38-point English–German gap that fell to four points when the sample was widened and nothing else changed. At the sample sizes common in this area, a large apparent language effect is an ordinary draw from a population with no effect at all.

4.4 Instability, and its ordering, replicate

Among prompt-platform cells observed three times, the share whose citation outcome was not unanimous was, in Study 1 and Study 2 respectively: Gemini 40.3% and 32.2%, Google AI Overviews 15.3% and 13.3%, ChatGPT 5.6% and 18.9%, Perplexity 1.4% and 2.2%. Gemini is the least stable and

Perplexity the most stable in both. ChatGPT is the one platform whose position moves between studies. The companion paper develops the consequences for sample-size planning.

4.5 Authenticity-seeking demand is not routed to rights holders

Study 2's prompts ask, in six languages, where to buy the genuine article, how to tell an original from a copy, and which retailers are authorised. Table 4 gives the most-cited domains.

#	Domain	Share of answers
1	vitra.com *(rights holder)*	13.6%
2	1stdibs.com	12.5%
3	whoppah.com	11.3%
4	**R1 (replica retailer)**	**9.7%**
5	mohd.it	9.1%
6	smow.com	6.8%
7	silvera.fr	6.4%
12	cassina.com *(rights holder)*	4.8%
14	hermanmiller.com *(rights holder)*	4.4%
15	flos.com *(rights holder)*	4.4%
17	louispoulsen.com *(rights holder)*	4.4%

Table 4. Most-cited domains on authenticity-seeking queries (Study 2, 1,080 answers). R1 is a replica retailer operated by the author, pseudonymised per §3.6.

Only one rights holder, Vitra, leads the table. Two of the top three cited domains are resale marketplaces. A replica retailer is fourth, ahead of Cassina, Herman Miller, Flos and Louis Poulsen — the companies that own the designs the query is asking about. On Google AI Overviews specifically, the rights holders were cited in 6.3% of answers, and in the German market in none at all.

We report this as a measurement, not as a claim about intent or liability. It says nothing about whether any retailer sought that placement, and we did not examine the content of the pages involved. What it does show is that a user asking an AI assistant where to buy the authentic article is, on this evidence, more likely to be pointed at a marketplace or a replica seller than at the manufacturer.

4.6 Citations are dispersed, not concentrated

Across Study 1's 28,371 citations we observe 5,755 distinct registrable domains. The ten most-cited account for 7.7% of all citations and the fifty most-cited for 14.7%; 1,328 domains are cited exactly once. The most frequently cited single domain is `reddit.com`, present in 10.7% of answers, followed by `youtube.com` at 7.7% — a statement about presence rather than dominance, since user-generated and marketplace domains together account for 4.5% of all citations. Generative answers in this setting do not funnel attention to a small set of incumbents.

5. Threats to validity

API versus consumer surfaces. Stated in §3.3; the central limitation.

Temporal validity. These systems change without notice. Our observations describe a window in September 2026, and every observation is timestamped so a future replication can be compared rather than merely contrasted.

Coverage. Six Western European languages, all well-resourced, is not the world. We make no claim about low-resource languages, where we would expect effects to be larger and where the question is more urgent.

Study 2 is smaller. 90 prompts against Study 1's 360. It is powered to replicate the platform ordering, which it does, and not to resolve fine differences between adjacent platforms or languages, which we accordingly do not claim.

Author interest. The author operates the replica retailers that are Study 1's focal brands and R1 in Study 2, and sells services in the measurement category this paper criticises. Study 2 exists because of that conflict: it repeats the central analysis on eight third-party rights holders with which the author has no relationship, and reproduces the result. The finding least convenient to the author's commercial interest — that the language gap his own earlier work claimed does not survive measurement — is the finding the paper leads with. The raw logs are released so this can be checked rather than trusted.

6. Discussion

The actionable disparity is across platforms, not across languages. A business absent from AI recommendations is, on this evidence, far more likely to be absent because of which assistant was asked than which language it was asked in. That inverts the advice the commercial category gives, and changes what a remedy looks like: not translating a catalogue, but understanding why one retrieval pipeline surfaces a site another does not. We do not identify that mechanism. The platforms differ in index, retrieval and ranking, all undisclosed, and our design cannot separate them.

Language may still matter where we could not measure it. A null result across six well-resourced European languages is not evidence about Urdu, Swahili or Vietnamese, where training and retrieval asymmetries are far larger. We regard that as the more urgent study.

Authenticity is a retrieval problem, not only a legal one. Section 4.5 suggests that the mapping from "I want the genuine article" to "here is the company that makes it" is weak in current generative search. Rights holders concerned with where authenticity-seeking demand lands may find that the binding constraint is retrieval behaviour rather than enforcement.

The category is selling a precision it does not have. Section 4.4 and the companion paper imply that a monitoring product sampling a few dozen prompts once per period cannot resolve the changes it reports. This is not an argument that measurement is worthless — the platform differences here are enormous and perfectly measurable. It is an argument that reported precision should match the sampling.

7. Reproducibility

The harness, prompt corpora, observation logs and analysis code are released under MIT (code) and CC BY 4.0 (data). A full replication requires API credentials for the four providers and costs approximately $58. The harness is resumable and logs every call, so a partial replication is possible at proportional cost. The companion methods paper is archived at doi:10.5281/zenodo.22480733.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. (2024). GEO: Generative Engine Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data*

Mining.

Allaham, M. and Diakopoulos, N. (2026). Synthetic Sources? Auditing Generative Search Engine Citations for Evidence of AI-Generated Sources. *arXiv preprint arXiv:2605.23684*.

Arshad, M. (2026). How Much Measurement Does an AI Visibility Claim Need? Replicate Instability and Minimum Detectable Effects in Generative Search Audits. *Zenodo*. doi:10.5281/zenodo.22480733.

Kumar, P. (2026). Generative Engine Optimization at Scale: Measuring Brand Visibility Across AI Search Engines. *arXiv preprint arXiv:2606.20065*.

Liu, N. F., Zhang, T., and Liang, P. (2023). Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., and Sandvig, C. (2021). Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends in Human–Computer Interaction*, 14(4).

Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Presented at *Data and Discrimination*, a preconference at the 64th Annual Meeting of the International Communication Association.